AD-A240 133

ENTATION PAGE

Form Approved
OMB No. 0704-0188

2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
Aug 91	Final 1 Feb 88 - 31 Jan 91

4. TITLE AND SUBTITLE
Visual Motion Perception (1-27 Atch)

5. FUNDING NUMBERS
G AFOSR-88-0140
PR 2313
TA A5
PE 61102F

6. AUTHOR(S)
Dr. George Sperling

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Dr. George Sperling, Human Information Processing Lab
N.Y.U. Dept of Psychology and Center for Neural Science
6 Washington Place, Room 930
New York, NY 10003

8. PERFORMING ORGANIZATION REPORT NUMBER
AFOSR-TR- 91 0757

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NL
Bldg 410
Bolling AFB
Washington, DC 20332-6448

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release;
distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The articles enclosed with this report describe work related to five aspects of visual information processing. (1) Continuing studies of two separate motion-computation systems in human vision and the derivation of the functional properties of each. (2) The investigation of 3D structure derived from 2D visual inputs. Herein we add to our previous evidence that structure from motion depends primarily on first-order motion computation, and we demonstrate restricted abilities of the second-order system. (3) A potent form of spatial contrast-gain-control was discovered and found to be not only frequency selective but also orientation specific. This form of local gain control may exemplify a universal form of neural normalization. (4) Studies of human pattern recognition of familiar shapes (such as letters) show that its statistical efficiency approaches an incredible 50% of the ideal detector's efficiency when the pattern is spatially bandpass filtered in a band whose wavelength is of the same order as the pattern itself (independent of the size of the retinal image) (5) Studies of real and simulated saccadic eye movements (in which the same sequence of images that is produced on the retina during saccadic eye movements is artificially produced on a stationary retina.) answer the following questions about human visual perception. (i) Why don't we see the smear produced on the retina during an eye movement? (ii) Why doesn't the world appear to move as a result of the image movements produced by eye movements? (iii) Does the visual system require sudden stimulus onsets (such as those produced by eye movements) to initiate processing episodes? (iv) To serve the perceptual construction of a stable representation of the world, is there a special memory to relate images produced by successive eye movements?

14. SUBJECT TERMS
vision, motion perception, ideal detectors, contrast gain-control

15. NUMBER OF PAGES

16. PRICE CODE

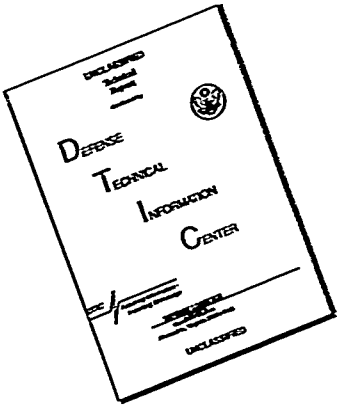
17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	unlimited

NSN 7540-01-280-5500

Standard Form 298 (Rev 2-89)
Prescribed by ANSI Std Z39-18

26 AUG 1991

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

Publications 1988-1991

1988 George Sperling and Thomas R. Riedl, Summation and masking between spatial frequency bands in dynamic natural visual stimuli. *Investigative Ophthalmology and Visual Science. ARVO Supplement*, 1988, *29, No. 3*, 139. (Abstract)

1988 Charles Chubb and George Sperling, Processing Stages in Non-Fourier Motion Perception. *Investigative Ophthalmology and Visual Science, ARVO Supplement*, 1988, *29, No. 3*, 266. (Abstract)

1988 Sperling, George. The magical number seven: Information processing then and now. In William Hirst (Ed), *The making of cognitive science: Essays in honor of George A. Miller*. Cambridge, UK: Cambridge University Press, 1988. Pp. 71-80.

1988 Riedl, Thomas R. and George Sperling. Spatial frequency bands in complex visual stimuli: American Sign Language. *Journal of the Optical Society of America A: Optics and Image Science*, 1988, *5*, 606-616.

1988 Chubb, Charles, and George Sperling. Drift-balanced random stimuli: A general basis for studying non-Fourier motion perception. *Journal of the Optical Society of America A: Optics and Image Science*, 1988, *5*, 1986-2006.

1988 *Sperling, George and Karl Gegenfurtner. Two transfer processes in iconic memory. *Bulletin of the Psychonomic Society*, 1988, *26*, 488. (Abstract)

1989 Chubb, Charles, and George Sperling. Second-order motion perception: Space-time separable mechanisms. *Proceedings: Workshop on Visual Motion*. (March 20-22, 1989, Irvine, California.) Washington, D.C: IEEE Computer Society Press, 1989. Pp. 126-138.

1989 Charles Chubb and George Sperling, Texture interactions determine apparent lightness. *Investigative Ophthalmology and Visual Science, ARVO Supplement*, 1989, *30, No. 3*, 161. (Abstract)

1989 George Sperling and Charles Chubo. Apparent motion derived from spatial texture. *Investigative Ophthalmology and Visual Science, ARVO Supplement*, 1989, *30, No. 3*, 425. (Abstract)

1989 Chubb, Charles, and George Sperling. Two motion perception mechanisms revealed by distance driven reversal of apparent motion. *Proceedings of the National Academy of Sciences, USA*, 1989, *86*, 2985-2989.

1989 Dosher, Barbara A., Michael S. Landy, and George Sperling. Ratings of kinetic depth in multi-dot displays. *Journal of Experimental Psychology. Human Perception and Performance*, 1989, *15*, 116-425.

1989 Sperling, George, Michael S. Landy, Barbara A. Dosher, and Mark E. Perkins The kinetic depth effect and the identification of shape. *Journal of Experimental Psychology. Human Perception and Performance*, 1989, *15*, 426-440.

1989 Barbara A. Dosher, Landy, Michael S., and George Sperling. Kinetic depth effect and optic flow 1. 3D shape from Fourier motion. *Vision Research*, 1989, *29*, 1789-1813.

1989 Sperling, George. Three stages and two systems of visual processing. *Spatial Vision*, 1989, *4 (Praxdny Memorial Issue)*, 183-207.

1989 Chubb, Charles, George Sperling, and Joshua A. Solomon. Texture interactions determine perceived contrast. *Proceedings of the National Academy of Sciences, USA*, 1989, *86*, 9631-9635

*Primary support is from AFOSR 88-0140 for all articles except as noted by *, which indicates primary support by ONR grant No N00014 88-K-0569

Atch. 1

1990 Sperling, George. Comparison of perception in the moving and stationary eye. In E. Kowler (Ed), *Eye Movements and their Role in Visual and Cognitive Processes*. Amsterdam, The Netherlands: Elsevier Biomedical Press, 1990. Pp. 307-351.

1990 George Sperling, Barbara A. Dosher, and Landy, Michael S., How to study the kinetic depth experimentally. *Journal of Experimental Psychology: Human Perception and Performance*, 1990, *16*, 445-450.

1990 Parish, David H., Sperling, George, and Landy, Michael, S. Intelligent temporal subsampling of American Sign Language using event boundaries. *Journal of Experimental Psychology: Human Perception and Performance*, 1990, *282-294*.

1990 Farrell, Joyce E., M. Pavel, and George Sperling. The visible persistence of stimuli in stroboscopic motion. *Vision Research*, 1990, *30*, 921-936.

1990 Sutter, Anne, George Sperling and Charles Chubb. Measuring the spatial frequency selectivity of second-order texture mechanisms. *Investigative Ophthalmology and Visual Science, ARVO Supplement*, 1990, *31, No. 4*, 104. (Abstract)

1990 Solomon, Joshua A, Charles Chubb, and George Sperling. The lateral inhibition of perceived textural contrast is orientation specific. *Investigative Ophthalmology and Visual Science, ARVO Supplement*, 1990, *31, No. 4*, 561. (Abstract)

1990 *Sperling, George and Weichselgartner, Erich. Episodic Theory of Visual Attention. *Bulletin of the Psychonomic Society*, 1990, *28*, 482. (Abstract)

1990 *Wurst, Stephen A., George Sperling, and Barbara Anne Dosher. Evidence for a central locus of short-term visual repetition memory. *Bulletin of the Psychonomic Society*, 1990, *28*, 514-515. (Abstract)

1991 Landy, Michael S., Barbara A. Dosher, George Sperling, and Mark E. Perkins. Kinetic depth effect and optic flow: 2. Fourier and non-Fourier motion. *Vision Research*, 1991, *31*, 859-876.

1991 Parish, David H. and George Sperling, Object spatial frequencies, retinal spatial frequencies, noise, and the efficiency of letter discrimination. Vision Research, *31*, 1399-1415.

1991 Solomon, Joshua A, and George Sperling. Can we see 2nd-order motion and texture in the periphery? *Investigative Ophthalmology and Visual Science*, ARVO Supplement, 1991, *32, No. 4*, 714. (Abstract)

1991 Werkhoven, Peter, Charles Chubb, and George Sperling. Texture-defined motion is ruled by an activity metric--not by similarity. *Investigative Ophthalmology and Visual Science*, ARVO Supplement, 1991, *32, No. 4*, 829. (Abstract)

1991 Sutter, Anne, George Sperling and Charles Chubb, Further measurements of the spatial frequency selectivity of second-order texture meachanisms. *Investigative Ophthalmology and Visual Science*, ARVO Supplement, 1991, *32, No. 4*, 1039. (Abstract)

1991 Chubb, Charles, and George Sperling. Texture quilts: Basic tools for studying motion-from-texture. *Journal of Mathematical Psychology*, 1991, *35*. (In press.)

*Primary support is from AFOSR 88-0140 for all articles except as noted by *, which indicates primary support by ONR grant No N00014 88 K-0569.

1991 *Wurst, S. A. and Sperling, G. Using repetition detection to define and localize the processes of selective attention. In D. E. Meyer and S. Kornblum (Eds.), *Attention and Performance XIV*, Erbaum: Hillsdale, NJ. (In press.)

Papers Under Submission for Publication, Technical Reports

1990 *Sperling, George, and Erich Weichselgartner. Episodic theory of the dynamics of spatial attention. *Mathematical Studies in Perception and Cognition, 89-8*, New York University, Department of Psychology, 1989. (Under revision, Psychological Review.)

1990 *Gegenfurtner, Karl and Sperling, George. Information transfer in iconic memory experiments. Mathematical Studies in Perception and Cognition, 90-8, New York University, Department of Psychology, 1990. (Submitted for publication.)

*Primary support is from AFOSR 88-0140 for all articles except as noted by *, which indicates primary support by ONR grant No. N00014 88 K-0569

AIR FORCE OF SCIENTIFIC RESEARCH (AFSC)
NOTICE OF TRANSMITTAL TO DTIC
This technical report has been reviewed and is
approved for public release IAW AFR 190-12
Distribution is unlimited.
Glo... Miller
STINFO Program Manager

Approved for public release;
distribution unlimited.

Texture Quilts: Basic Tools For Studying Motion-From-Texture

Charles Chubb

Department of Psychology, Rutgers University

New Brunswick, NJ 08903

George Sperling

Human Information Processing Laboratory

Psychology Department and Center for Neural Sciences,

New York University, 6 Washington Place

New York, NY 10003

Address correspondence to:

Charles Chubb
Department of Psychology
Rutgers University
New Brunswick, NJ 08903

201 932-2640

Atch-2

Abstract

A theoretical foundation and concrete stimulus-construction methods are provided for studying motion-from-spatial-texture without contamination by motion mechanisms sensitive to other aspects of the signal. Specifically, examples are constructed of a special class of random stimuli called *texture quilts*. Although, as we demonstrate experimentally, certain texture quilts display consistent apparent motion, it is proven that their motion content (a) is unavailable to standard motion analysis (such as might be accomplished by an Adelson/Bergen motion-energy analyzer, a Watson/Ahumada motion sensor, or by any elaborated Reichardt detector), and (b) cannot be exposed to standard motion analysis by any purely temporal signal transformation no matter how nonlinear (e g., temporal differentiation followed by rectification). Applying such a purely temporal transformation to any texture quilt produces a spatiotemporal function P whose motion is unavailable to standard motion analysis: The expected response of every Reichardt detector to P is 0 at every instant in time. The simplest mechanism sufficient to sense the motion exhibited by texture quilts consists of three successive stages: (i) a purely spatial linear filter (ii) a rectifier (but not a perfect square law) to transform regions of large negative or positive responses into regions of high positive values, and (iii) standard motion analysis.

1. Introduction.

Standard motion analysis. The extensive literature on the motion of random-dot cinematograms (Anstis, 1970; Julesz, 1971; Braddick, 1973, 1974; Lappin & Bell, 1976; Bell & Lappin, 1979; Baker & Braddick, 1982a, 1982b; Chang & Julesz, 1983a, 1983b, 1985; Ramachandran & Anstis, 1983; Nakayama & Silverman, 1984; van Doorn & Koenderink, 1984) points toward the view that a "short-range" system (Braddick, 1973, 1974) submits the raw spatiotemporal luminance function directly to *standard motion analysis* (such as might be accomplished by an Adelson/Bergen motion-energy detector (Adelson & Bergen, 1985), a Watson/Ahumada motion sensor (Watson & Ahumada, 1983a, 1983b, 1985), an elaborated Reichardt detector (van Santen & Sperling, 1984, 1985), or some variants of a gradient detector (Marr & Ullman, 1981; Adelson & Bergen, 1986)).

Fourier and nonFourier mechanisms. An impressive number of observations suggests that standard motion analysis is not the whole story (Sperling, 1976; Ramachandran, Rao & Vidyasagar, 1973; Petersik, Hicks & Pantle, 1978; Ramachandran, Ginsburg & Anstis, 1983; Lelkins & Koenderink, 1984; Derrington & Badcock, 1985; Green, 1986; Pantle & Turano, 1986; Derrington & Henning, 1987; Turano & Pantle, 1988; Bowne, McKee & Glaser, 1989; Cavanagh, Arguin & von Grunau, 1989). In particular, Chubb and Sperling (1987, 1988) have demonstrated a variety of stimuli that display consistent, unambiguous apparent motion, yet that do not systematically stimulate mechanisms that apply standard motion analysis directly to luminance. For reasons that will become clear in Section 2, we call any motion system that applies standard analysis to the raw signal as a *Fourier* mechanism, and we refer to any system that applies standard analysis to a nonlinear transformation of the signal as a *nonFourier* mechanism.

Microbalanced stimuli. The methods used by Chubb & Sperling to construct stimuli whose obvious and consistent motion content cannot be revealed by applying standard motion analysis directly to luminance are founded on the notion of a *microbalanced* random stimulus. In Section 2.3.5, we show that the expected response of any standard motion analyzer applied directly to any microbalanced random stimulus is equal to the expected response of the corresponding analyzer tuned

te motion of the same type, but in the opposite direction.

Microbalanced random stimuli allow us to differentially stimulate nonFourier motion mechanisms without systematically engaging Fourier mechanisms. This is the source of their importance in the study of motion perception.

There are probably several types of nonFourier motion mechanisms, distinguished by the different nonlinear transformations they apply to the signal prior to standard motion analysis. In this paper, we extend the theory of microbalanced random stimuli in order to develop methods for constructing stimuli that selectively engage specific classes of nonFourier mechanisms without stimulating either Fourier mechanisms or other classes of nonFourier mechanisms.

Pointwise transformations, static nonlinearities. A transformation T is called *pointwise* if the output of T at any point (x,y,t) in space-time depends only on the (stimulus) input value at that point. A *nonlinear* pointwise transformation sometimes is called a *static nonlinearity*. For instance, simple rectifiers and thresholders are pointwise transformations. In Section 3, we address the problem of isolating the class of nonFourier mechanisms that apply a simple pointwise transformation prior to standard motion analysis from the class of all those mechanisms that apply more complicated transformations. The central result in this Section is proposition 3.2 which provides necessary and sufficient conditions for a random stimulus I to be such that any pointwise transformation of I is microbalanced.

Purely temporal transformations and texture quilts. The results with pointwise transformations are extended in Section 4 to purely temporal transformations (defined in Section 2.2). Whereas, for a pointwise transformation, the transformed value at the point (x,y,t) depends only on the stimulus value at (x,y,t), in a purely temporal transformation the transformed value at (x,y,t) may depend in any way whatsoever on the entire history of stimulus values at (x,y). We define the class of stimuli called *texture quilts* (Definition 4.1) whose importance derives from the fact (proven in proposition 4.3) that any purely temporal transformation of a texture quilt is microbalanced. Concrete methods

are provided for constructing *binary* and *sinusoidal* texture quilts that display consistent motion.

In Section 5, these construction methods are applied in an experiment designed to demonstrate the effectiveness of three textural properties as carriers of motion information. The textural properties are (i) spatial frequency variation (ii) orientation variation, and (iii) variation between perceptually distinct textures with identical expected energy spectra.

2. Preliminaries.

This section states the background facts presupposed by the main discussion of the paper.

2.1. Discrete dynamic visual stimuli.

Notation. Let $\mathbf{R}$ denote the real numbers, and $\mathbf{Z}$ ($\mathbf{Z}^+$) the integers (positive integers). We use square brackets to enclose arguments of discrete functions, and parentheses to enclose arguments of continuous functions.

The range of a stimulus. We want the term "stimulus" to refer not only to the luminance function submitted as input to the retina, but to any physiologically reasonable transformation of the spatiotemporal luminance function which might be submitted as input to a component processor of the visual system. Consequently, although luminance is physically a non-negative quantity, we do not apply this constraint to the class of functions we admit as stimuli. We allow stimuli to take values throughout the positive and negative real numbers.

The domain of a stimulus. To remain close to our intuitions about neurally realized visual processors, we take stimuli to be a functions of the discrete domain $\mathbf{Z}^3$ (where the dimensions correspond to horizontal and vertical space, and time). In addition, for mathematical convenience, and without loss of physiological plausibility, we require a stimulus to be 0 almost everywhere in its (infinite) domain.

The definition of a stimulus. We call any function $I:\mathbf{Z}^3 \rightarrow \mathbf{R}$ a *stimulus* provided $I[x,y,t] = 0$ for all but finitely many points of $\mathbf{Z}^3$.

We shall be considering stimuli as functions of two spatial dimensions x, y and time t.

Stimulus contrast. As is now well-established (e.g., Shapley & Enroth-Cugell, 1984), early retinal gain-control mechanisms pass not stimulus luminance, but rather a signal approximating stimulus *contrast*, the normalized deviation at each time t of luminance at each point (x,y) in the visual field from a "background level", or "level of adaptation", which reflects the average luminance over points proximal to (x,y,t) in space and time. Because the transformation from luminance to contrast is a processing stage that is general to all of vision. we shall drop reference to mean luminance L_o, and characterize L only by its *contrast modulation function*, C:

$$C = \frac{L}{L_o} - 1. \tag{1}$$

What we shall argue in this paper is that the broad-band spatial filtering that mediates the step from luminance to contrast is succeeded by additional filtering stages in which a number of *narrowly tuned* spatial filters are applied to the visual signal, their output rectified, and the resulting spatiotemporal signal processed for motion information.

The history of a stimulus at a point in space. For any stimulus I, any point $(x, y) \in \mathbf{Z}^2$, we define $I_{(x,y)}$, the *history of* I *at* (x, y), by setting

$$I_{(x,y)}[t] = I[x, y, t] \tag{2}$$

for all $t \in \mathbf{Z}$.

Space-time separable stimuli. A stimulus I is called *space-time* separable iff I can be expressed as the product of a spatial function $f:\mathbf{Z}^2 \rightarrow \mathbf{R}$ and a temporal function $g:\mathbf{Z} \rightarrow \mathbf{R}$: For all $(x, y, t) \in \mathbf{Z}^3, I[x, y, t] = f[x, y]g[t]$.

The Fourier transform of a stimulus. Because any stimulus I is nonzero at only a finite number of points, the energy in I is finite, implying that I has a well-defined Fourier transform.

We denote I's Fourier transform by $\bar{I}$; writing j for the complex number $(0, 1)$,

$$\bar{I}(\omega,\theta,\tau) = \sum_{x=-\infty}^{\infty}\sum_{y=-\infty}^{\infty}\sum_{t=-\infty}^{\infty} I[x,y,t]e^{-j(\omega x+\theta y+\tau t)}. \qquad (3.$$

Although $\bar{I}$ is defined for all real numbers ω,θ,τ, it is periodic over 2π in each argument. This fact is reflected in the inverse transform:

$$I[x,y,t] = \frac{1}{(2\pi)^3}\int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi} \bar{I}(\omega,\theta,\tau)\, e^{j(\omega x+\theta y+\tau t)}\, d\omega\, d\theta\, d\tau. \qquad (4.$$

In the Fourier domain, we consistently use ω to index frequencies relative to x, θ frequencies relative to y, and τ frequencies relative to t.

The function $\mathbf{0}$. We write $\mathbf{0}$ for any function that assigns 0 to each element in its domain. Thus, $\mathbf{0}$ defined on Z^3 is the stimulus that is zero throughout space and time. We also write $\mathbf{0}$ for the temporal function that sets $\mathbf{0}[t] = 0$ for all $t \in Z$.

2.2. Mappings and stimulus transformations.

Let Ω be the set of all real-valued functions of Z^3, and call any function of Ω into Ω a *mapping*. (We shall need the general notion of a mapping only briefly in order to specify the subset of well-behaved mappings called transformations.) For any mapping M and any $I \in \Omega$, $M(I)$ is a real-valued function of Z^3; accordingly, we write $M(I)[x,y,t]$ for the value of $M(I)$ at any point $(x,y,t) \in Z^3$.

If it is continuous, a function $f:\mathbb{R} \to \mathbb{R}$ submits to a wide range of useful operations. For instance, if f is continuous, it can be integrated over any finite interval. Of course, f need not be continuous to meet this condition. For instance, f is integrable over any finite interval if f is discontinuous at only a finite number of points in any finite interval. If f is integrable over any finite interval, and if f also is bounded, then for any function g for which $\int_{\mathbb{R}} g$ converges, $\int_{\mathbb{R}} fg$ also converges. In particular, $\int_{\mathbb{R}} fg$ converges if g is a density function. For the results reported here, we restrict our attention to a special class of mappings, which we shall call stimulus transformations, that have

properties analogous to those of the well-behaved function f. We specify these desirable properties in the following paragraph.

Continuous mappings; finitely integrable mappings; bounded mappings. For any $I \in \Omega$, any $p \in \mathbf{R}$, any $\psi \in \mathbf{Z}^3$, we write $I_{\psi \leftarrow p}$ for the element of Ω that is identical to I at all locations of $\mathbf{Z}^3$ except ψ, where it takes the value p. Any mapping M is called *continuous* if $M(I_{\psi \leftarrow p})[\zeta]$ is a continuous function of p for any $I \in \Omega$, and any $\psi, \zeta \in \mathbf{Z}^3$. M is called *finitely integrable* if, for any such I, ψ, and ζ, $M(I_{\psi \leftarrow p})[\zeta]$ is an integrable function of p over any finite interval. Finally, M is called *bounded* if, for any such I, ψ, and ζ, $M(I_{\psi \leftarrow p})[\zeta]$ is is a bounded function of p over the set of real numbers.

The definition of a stimulus transformation. A *stimulus transformation* (which we shall often refer to simply as a *transformation*) is a bounded, finitely integrable, mapping T such that $T(S)$ is a stimulus for any stimulus S, and $T(0) = 0$.

There are other reasonable constraints we might impose on the notion of a stimulus transformation. For instance, we might require a stimulus transformation to be time-invariant and causal. However, we do not include these conditions in our definition because they are not required for the results we report.

Purely temporal stimulus transformations. Let Ω_T be the set of all functions mapping $\mathbf{Z}$ into $\mathbf{R}$. A transformation H is called *purely temporal* iff there exists a function $H_T: \Omega_T \rightarrow \Omega_T$ such that for any stimulus I, any $(x, y, t) \in \mathbf{Z}^3$,

$$H(I)[x, y, t] = H_T(I_{(x,y)})[t]. \tag{5}$$

That is, the value at the point $(x, y, t) \in \mathbf{Z}^3$ that results from applying H to I depends only on the history of I at (x, y). Since it is obvious from the context, we drop the distinction between H and H_T, and allow H to be applied both to full-fledged stimuli and to simple functions of time. Thus, for any temporal function $P: \mathbf{Z} \rightarrow \mathbf{R}$, we shall write $H(P)$ to indicate the temporal function $H_T(P)$.

We shall be particularly concerned with two types of transformations: *pointwise* transformations and *linear, shift-invariant* transformations.

Pointwise transformations and rectifiers. For any functions $f: A \rightarrow B$ and $g: B \rightarrow C$, the *composition* $g \bullet f: A \rightarrow C$ is given by

$$g \bullet f(a) = g(f(a)) \tag{6}$$

for any $a \in A$. For any $f: \mathbf{R} \rightarrow \mathbf{R}$, we call the mapping $f\bullet$, yielding the spatiotemporal function $f \bullet I$ when applied to stimulus I, a *pointwise* mapping (because its output value at any point in space-time depends only on its input value at that point).

As is evident, $f\bullet$ is a transformation iff (i) $f(0) = 0$, (ii) f is bounded on $\mathbf{R}$, and (iii) f is integrable over any bounded real interval. A transformation $f\bullet$ is called a *positive half-wave rectifier* if f is monotonically increasing, and $f[v] = 0$ for all $v \leq 0$; $f\bullet$ is called a *negative half-wave rectifier* if f is monotonically decreasing, and $f[v] = 0$ for $v \geq 0$. Finally, $f\bullet$ is called a *full-wave* rectifier if f is a monotonically increasing function of absolute value.

Linear, shift-invariant (LSI) transformations. For any offset $\psi \in \mathbf{Z}^3$, define the mapping S^{ψ} by

$$S^{\psi}(I)[\zeta] = I[\zeta - \psi] \tag{7}$$

for any $I \in \Omega$. Thus $S^{\psi}(I)$ is derived by shifting I by the offset ψ in $\mathbf{Z}^3$. Any mapping M is called *shift-invariant* iff

$$S^{\psi}(M(I)) = M(S^{\psi}(I)) \tag{8}$$

for any $\psi \in \mathbf{Z}^3$, any $I \in \Omega$. In addition, M is *linear* iff for any $I, J \in \Omega$, any real numbers κ and λ

$$M(\kappa I + \lambda J) = \kappa M(I) + \lambda M(J). \tag{9}$$

As is well known, any linear, shift-invariant (*LSI*) transformation can be expressed as a *convolution*, which is defined for any $u \in \mathbf{Z}^3$ by

$$(k * I)[u] = \sum_{v \in \mathbf{Z}^3} k[u - v]\, I[v], \tag{10}$$

for some $k: \mathbf{Z}^3 \rightarrow \mathbf{R}$. The function k is called the *impulse response* of the transformation $k*$.

2.3. Random stimuli.

For any real random variable X with density f, we write $E[X]$ for the *expectation of* X:

$$E[X] = \int_{\mathbf{R}} x f(x)\, dx. \tag{11}$$

The notion of a random stimulus generalizes that of a (non-random) stimulus in that the values assigned points in space-time by a random stimulus are random variables (with finite variances) rather than constants.

The definition of a random stimulus. Call any family $\{R[x,y,t] \parallel (x,y,t) \in \mathbf{Z}^3\}$ of jointly distributed random variables a *random stimulus* provided

(i) $R[x,y,t]$ is constant and equal to 0 for all but finitely many $(x,y,t) \in \mathbf{Z}^3$,

and

(ii) $E\left[R[x,y,t]^2\right]$ exists for all $(x,y,t) \in \mathbf{Z}^3$.

As with non-random stimuli, we write $\bar{R}$ for the Fourier transform of any random stimulus R; and, for any $\chi = (x,y) \in \mathbf{Z}^2$ we write R_χ for the temporal random function defined by

$$R_\chi[t] = R[\chi, t] \tag{12}$$

for all times $t \in \mathbf{Z}$.

Space-time separable random stimuli. We call a random stimulus R *space-time separable* iff R is space-time separable with probability 1.

Constant stimuli. Any ordinary stimulus can be regarded as a random stimulus that does not vary across independent realizations. We call such such unvarying stimuli *constant*.

The motion-from-Fourier-components principle. *Parseval's relation* states that the energy in a stimulus is proportional to the energy in its Fourier transform. Individual spatiotemporal Fourier components are drifting sinusoidal gratings. Thus, we can add up the energy in a dynamic visual stimulus either point-by-point in space-time, or drifting sinusoid by drifting sinusoid. A commonly

encountered rule of thumb (Watson, Ahumada & Farrell, 1986; Watson & Ahumada, 1983b; van Santen & Sperling, 1985) for predicting the apparent motion of an arbitrary stimulus $I[x,y,t]=f[x,t]$ (constant in the vertical dimension of space), is the *motion-from-Fourier-components* principle: For I regarded as a linear combination of drifting sinusoidal gratings, if most of I's energy is contributed by rightward-drifting gratings, then perceived motion should be to the right. If most of the energy resides in the leftward-drifting gratings, perceived motion should be to the left. Otherwise I should manifest no decisive motion in either direction.

Drift-balanced random stimuli. The class of *drift-balanced* random stimuli (Chubb & Sperling, 1987, 1988) provides a rich pool of counterexamples to the motion-from-Fourier-components principle. A random stimulus R is drift balanced iff the expected energy in R of each drifting sinusoidal component is equal to the expected energy of the component of the same spatial frequency, drifting at the same rate, but in the opposite direction. The term *drift balanced* is defined formally as follows.

Definition of a drift-balanced random stimulus. Call any random stimulus R *drift balanced* iff

$$E\left[|\bar{R}(\omega,\theta,\tau)|^2\right] = E\left[|\bar{R}(\omega,\theta,-\tau)|^2\right] \tag{13}$$

for all $(\omega,\theta,\tau) \in \mathbf{R}^3$.[1]

Thus, for any class of spatiotemporal linear receptors tuned to stimulus energy in a certain spatiotemporal frequency band, a drift-balanced random stimulus will, on the average, stimulate equally well those receptors tuned to the corresponding band of opposite temporal orientation.

Microbalanced random stimuli. Consider the following two-flash stimulus S: In flash 1, a bright spot (call it Spot 1) appears. In flash 2, Spot 1 disappears, and two new spots appear, one to the left and one symmetrically to the right of Spot 1. As one might suppose, S is drift balanced. On the

[1]. For a proof that the expected energy of the Fourier transform of any random stimulus is everywhere well-defined see Chubb & Sperling, 1988, appendix A

other hand, it is equally clear that a Fourier motion detector whose spatial reach encompasses the location of Spot 1 and only *one* of the Spots in flash 2 may well be stimulated in a fixed direction by S. Thus, although S is drift balanced, some Fourier motion detectors may be stimulated strongly and systematically by S. These detectors can be differentially selected by *spatial windowing*, and thereby the drift-balanced stimulus S is converted into a non-drift-balanced stimulus by multiplying it by an appropriate space-time separable function. The following subclass of drift-balanced random stimuli cannot be made non-drift-balanced by space-time separable windowing.

Definition of a microbalanced random stimulus. Call any random stimulus I *microbalanced* iff the product WI is drift balanced for any space-time separable function W.

One can think of the multiplying function W as a "window" through which a spatiotemporal subregion of I can be "viewed" in isolation. The space-time separability of W insures that W is "transparent" with respect to the motion-content of the region to which it is applied: W does not distort I's motion with any motion content of its own. The fact that I is microbalanced means that any subregion of I encountered through a "motion-transparent window" is drift balanced.

The following characterization of the class of microbalanced random stimuli, and all other results stated without proof in this section are from Chubb and Sperling (1988).

2.3.1. *A random stimulus I is microbalanced if and only if*

$$E\Big[I[x,y,t]I[x',y',t'] - I[x,y,t']I[x',y',t]\Big] = 0 \qquad (14)$$

for all $x,y,t,x',y',t' \in \mathbf{Z}$.

Some other relevant facts about microbalanced random stimuli:

2.3.2. *For any independent microbalanced random stimuli I and J,*

I. *the product IJ is microbalanced,*

and

II. *the convolution $I * J$ is microbalanced.*

2.3.3. *(a) Any space-time separable random stimulus is microbalanced; (b) any constant microbalanced stimulus is space-time separable.*

The following result is useful in constructing a wide range of microbalanced random stimuli which display striking apparent motion.

2.3.4. *Let Γ be a family of pairwise independent, microbalanced random stimuli, all but at most one of which have expectation* **0**. *Then any linear combination of Γ is microbalanced.*

Reichardt detectors and microbalanced random stimuli. Two Fourier motion detectors proposed for psychophysical data (Adelson & Bergen, 1985; Watson & Ahumada, 1983a, 1983b) can be recast as *Reichardt detectors* (Adelson & Bergen, 1985; van Santen & Sperling, 1985). The Reichardt detector has many useful properties as a motion detector without regard to its specific instantiation (van Santen & Sperling, 1984, 1985).

FIG 1

Figure 1 shows a diagram of the Reichardt detector. It consists of spatial receptors characterized by spatial functions f_1 and f_2, temporal filters g_{1*} and g_{2*}, multipliers, a differencer, and another temporal filter h_*. The spatial receptors f_i, $i = 1, 2$, act on the input stimulus I to produce intermediate outputs,

$$y_i[t] = \sum_{(x,y)\in Z^2} f_i[x,y]\, I[x,y,t]. \tag{15}$$

At the next stage, each temporal filter g_{j*} transforms its input y_i $(i, j = 1, 2)$, yielding four temporal output functions: $g_j * y_i$. The left and right multipliers then compute the products

$$\Big[y_1 * g_1[t]\Big]\Big[y_2 * g_2[t]\Big] \quad \text{and} \quad \Big[y_1 * g_2[t]\Big]\Big[y_2 * g_1[t]\Big] \quad \text{respectively}, \tag{16}$$

and the differencer subtracts the output from the right multiplier from that of the left multiplier:

$$D[t] = \Big[y_1 * g_1[t]\Big]\Big[y_2 * g_2[t]\Big] - \Big[y_1 * g_2[t]\Big]\Big[y_2 * g_1[t]\Big]. \tag{17}$$

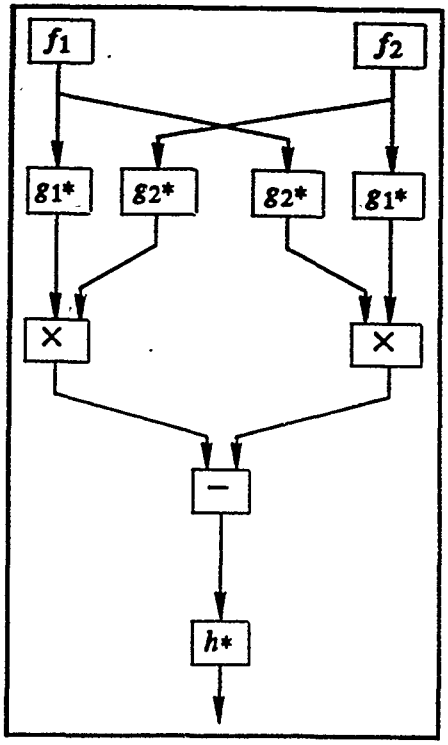


Fig. 1. The Reichardt detector. Let I be a random stimulus. Then, in response to I, for $i = 1, 2$, the box containing the spatial function $f_i: \mathbf{Z}^2 \to \mathbf{R}$, outputs the temporal function, $\sum_{(x,y)\in\mathbf{Z}^2} f_i[x,y]\,I[x,y,t]$; each of the boxes marked g_i* outputs the convolution of its input with the temporal function $g_i: \mathbf{Z} \to \mathbf{R}$; each of the boxes marked with a multiplication sign outputs the product of its inputs, the box marked with a minus sign outputs its left input minus its right, and the box containing $h*$ outputs the convolution of its input with the temporal function $h: \mathbf{Z} \to \mathbf{R}$. To see how the Reichardt detector senses motion, suppose f_2 is identical to f_1, but shifted in space by some offset, and suppose the filters g_1* do not alter their input, while the filters g_2* simply delay their input by some amount δ_t of time. Then a rigidly translating pattern moving in the direction of box f_2's offset from box f_1 will elicit some time-varying response from box f_1, and the same response a short time later from box f_2. If that "short time later" is precisely δ_t, the output of the righthand multiplier will be positive as long as the pattern keeps drifting. This will result in a net nega tive Reichardt detector output. If the pattern drift is in the opposite direction, the detector response will be positive.

The final output is produced by applying the filter $h*$, whose purpose is to smooth the time-varying, differencer output D. Since many Fourier mechanisms can be expressed as, or closely approximated by, Reichardt detectors (van Santen & Sperling, 1985; Adelson & Bergen, 1985, 1986), the following characterization of the class of microbalanced stimuli can be regarded as the cornerstone of the claim that microbalanced random stimuli bypass Fourier motion mechanisms.

2.3.5. *For any random stimulus I, the following conditions are equivalent:*

1. I is microbalanced.

2. the expected response of every Reichardt detector to I is 0 at every instant in time

Proof. Chubb & Sperling (1988) proved that I implies II. To obtain the reverse implication, note that if II holds, then, in particular, for any points $(x,y),(x',y') \in \mathbf{Z}^2$ and any $\delta_t \in \mathbf{Z}$, the expected response to I is the temporal function 0 for a particular simple Reichardt detector that computes

$$I[x,y,t]I[x',y',t-\delta_t] - I[x,y,t-\delta_t]I[x',y',t]. \tag{18}$$

This Reichardt detector is constructed by making (i) f_1 (of Fig. 1) the function that takes the value 1 at (x,y) and 0 everywhere else, (ii) f_2 the function that takes the value 1 at (x',y') and 0 everywhere else, (iii) each of g_1* and $h*$ the identity transformation, and (iv) g_2* the filter that delays its input by δ_t units of time. However, if the expected response to I is 0 throughout time for any such Reichardt detector, then Eq. (14) holds, and proposition 2.3.1 implies that I is microbalanced. ∎

3. Random stimuli microbalanced under all pointwise transformations.

The main purpose of this paper is to provide tools for differentially stimulating specific types of nonFourier motion mechanisms without engaging either Fourier mechanisms or other types of non-Fourier mechanisms. A nonFourier motion mechanism is one that applies an initial nonlinear transformation to the visual signal and subjects the output to standard motion analysis. In this section, we provide some results relevant to the psychophysical problem of stimulating nonFourier mechan-

isms whose initial transformation is nonpointwise without engaging any mechanism whose initial transformation is pointwise. The main finding is stated in proposition 3.2, which provides necessary and sufficient conditions for a random stimulus I to be such that $f \bullet I$ is microbalanced for any pointwise transformation $f \bullet$. In Section 4 we shall apply this result to construct random stimuli (texture quilts) which are microbalanced, and are, moreover, guaranteed to remain microbalanced after any purely temporal transformation. Such stimuli are useful for selectively stimulating nonFourier motion mechanisms that extract motion information from stimuli that have undergone nonlinear *spatial* stimulus transformations.

We begin by considering an example of a stimulus (Chubb & Sperling, 1987, 1988) that is microbalanced under all pointwise transformations, but whose motion can be revealed by a purely temporal nonlinear transformation.

3.1. Stimulus J: Traveling reversal of a random black-or-white vertical bar pattern. Let $M \in Z^+$. We construct the random stimulus J of $M+1$ frames indexed $0, 1, ..., M$, each of which contains M vertical bars, indexed $1, 2..., M$ from left to right. In frame 0 of stimulus J, all M vertical bars first appear. The contrast of each bar is 1 or -1 with equal probability, and bar contrasts are jointly independent. In each successive frame m, $m = 1, 2, ..., M$, the m^{th} rectangle flips its contrast to 1 if its previous contrast was -1; otherwise it flips from 1 to -1. In frame 1, rectangle 1 flips contrast; in frame 2, rectangle 2 flips, and in successive frames, successive rectangles flip contrast from left to right, until the Mth rectangle flips in frame M, after which all the rectangles turn off. An xt cross-section of frames 0 to M of J is shown in Fig. 2a.

FIG 2

The traveling contrast-reversal, stimulus J, is easily expressed as a sum of pairwise independent, space-time separable random stimuli, all with expectation 0; thus propositions 2.3.3a and 2.3.4 imply that J is microbalanced. Moreover, it is easy to see that, because J's frames are comprised of

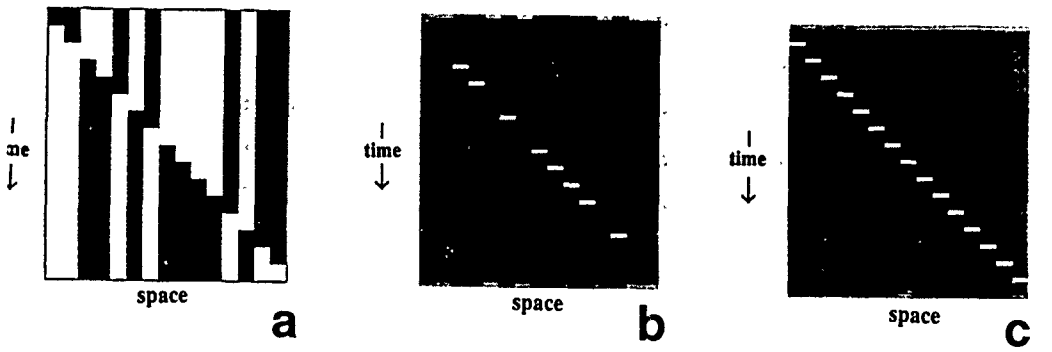


Fig. 2. Exposing the motion of the traveling contrast-reversal of the random black-or-white vertical bar pattern J to standard motion-analysis. (a) An xt cross-section of J. (b) An xt cross-section of the partial derivative of J with respect to time. (c) An xt cross-section of $|\partial J/\partial t|$. Each of J and $\partial J/\partial t$ is microbalanced. However, $|\partial J/\partial t|$ is not. In particular, $|\partial J/\partial t|$ has most of its energy at those frequencies whose velocity is equal to the velocity of the traveling contrast-reversal.

only two values, any pointwise transformation of J merely serves to rescale each of J's frames, and to shift it by a constant: that is, for any $f:\mathbf{R}\rightarrow\mathbf{R}$, $f\bullet J=\lambda J+K$, where $\lambda\in\mathbf{R}$, and K is a stimulus that assigns a constant value across all points at which J is nonzero. Clearly, $f\bullet J$ is another microbalanced random function (This follows easily from proposition 2.3.4). Thus, pointwise transformations fail to expose J's motion.

Exposing J's motion to standard analysis. Perhaps the simplest way to extract J's motion is to full-wave rectify the partial derivative of J taken with respect to time. The stages of this transformation are illustrated in Figs. 2b and 2c. Fig. 2b shows $\partial J/\partial t$. This function is itself microbalanced (propositions 2.3.2 II. and 2.3.3a imply that any purely temporal LSI transformation of a microbalanced random stimulus is microbalanced). However, $|\partial J/\partial t|$ (Fig. 2c) has most of its energy at those spatiotemporal frequencies whose velocity is equal to the velocity of the traveling contrast-reversal whose motion we wish to detect. Thus we see that, although J's motion cannot be exposed to standard analysis by a simple pointwise transformation, a temporal linear filter followed by a pointwise nonlinearity does suffice.

We turn now to the problem of stipulating the general conditions that a random stimulus I must satisfy so that $f\bullet I$ will be microbalanced for any pointwise transformation $f\bullet$. Call any random stimulus I *microbalanced under* a given transformation T iff $T(I)$ is microbalanced.

We state the following basic proposition (3.2) and its subsequent corollary (3.3) for continuously distributed random stimuli. The corresponding result for discretely distributed random stimuli is simpler and should be evident.

3.2. Necessary and sufficient conditions for a random stimulus to be microbalanced under all pointwise transformations. *Let I be a random stimulus such that for any $(x,y,t),(x',y',t')\in\mathbf{Z}^3$, $(I[x,y,t],I[x',y',t'])$ has a continuous joint density. Then the following conditions are equivalent:*

1 I is microbalanced under all pointwise transformations

2. *For all* $x, y, t, x', y', t' \in \mathbf{Z}$, *the joint density* f *of* $(I[x, y, t], I[x', y', t'])$ *and the joint density* g *of* $(I[x, y, t'], I[x', y', t])$ *satisfy*

$$f(p, q) + f(q, p) = g(p, q) + g(q, p) \tag{19}$$

for any $p, q \in \mathbf{R}$ *such that* $p \neq 0$ *and* $q \neq 0$.

Proof. Set $\kappa = I[x, y, t]$, $\lambda = I[x', y', t']$, $\gamma = I[x, y, t']$, and $\nu = I[x', y', t]$. Thus, (κ, λ) is distributed in $\mathbf{R}^2$ with density f and (γ, ν) is distributed with density g.

(2. implies 1.): By definition of any pointwise transformation $h\bullet$, we have $h(0) = 0$. Thus we need integrate only over values of κ and λ which are both nonzero in computing the expectation $E[h(\kappa)\, h(\lambda)]$. In particular, if Eq. (19) is satisfied for all $p \neq 0$ and $q \neq 0$, then $h \bullet I$ is microbalanced since

$$\begin{aligned} E[h(\kappa)\, h(\lambda)] &= \frac{1}{2}\left[\int_{\mathbf{R}}\int_{\mathbf{R}} h(p)\, h(q)\, f(p, q)\, dp\, dq + \int_{\mathbf{R}}\int_{\mathbf{R}} h(q)\, h(p)\, f(q, p)\, dq\, dp\right] \\ &= \frac{1}{2}\left[\int_{\mathbf{R}}\int_{\mathbf{R}} h(p)\, h(q)\, f(p, q)\, dp\, dq + \int_{\mathbf{R}}\int_{\mathbf{R}} h(p)\, h(q)\, f(q, p)\, dp\, dq\right] \\ &= \frac{1}{2}\int_{\mathbf{R}}\int_{\mathbf{R}} h(p)\, h(q)\, (f(p, q) + f(q, p))\, dp\, dq \\ &= \frac{1}{2}\int_{\mathbf{R}}\int_{\mathbf{R}} h(p)\, h(q)\, (g(p, q) + g(q, p))\, dp\, dq = E[h(\gamma)\, h(\nu)]. \end{aligned} \tag{20}$$

(Note: the boundedness & finite integrability of $h\bullet$ ensure that these expectations exist.)

(Not 2. implies not 1.): On the other hand, suppose Eq. (19) fails for some $x, y, t, x', y', t' \in \mathbf{Z}$. One way in which this might happen is if $f(r, r) > g(r, r)$ for some nonzero $r \in \mathbf{R}$. In this case, there exists a neighborhood N of r, not including 0, such that $f(m, n) > g(m, n)$ for all $m, n \in N$. Thus, for the function $h: \mathbf{R} \to \mathbf{R}$ defined by

$$h(n) = \begin{cases} 1 \text{ if } n \in N, \\ 0 \text{ otherwise}, \end{cases} \tag{21}$$

$h\bullet$ is a pointwise transformation (the function h is bounded on $\mathbf{R}$, finitely integrable, and $h(0) = 0$). However, $h \bullet I$ is not microbalanced since

$$E[h(\kappa)\,h(\lambda)] = \iint_{NN} f(m,n)\,dm\,dn \;>\; \iint_{NN} g(m,n)\,dm\,dn \;=\; E[h(\gamma)\,h(\nu)]. \tag{22}$$

To recapitulate, if Condition 2 fails because there exists a nonzero $r \in \mathbf{R}$ for which $f(r,r) \neq g(r,r)$, then Condition 1 fails (I is not microbalanced under all pointwise transformations).

The only other way in which Condition 2 can fail is if $f(r,r) = g(r,r)$ for all $r \neq 0$ in $\mathbf{R}$, but for some $p, q \in \mathbf{R}$, with neither p nor q equal to 0, $f(p,q) + f(q,p) > g(p,q) + g(q,p)$. In this case, we obtain disjoint neighborhoods M of p and N of q, neither including 0, such that

$$f(m,n) + f(n,m) > g(m,n) + g(n,m) \tag{23}$$

for all $m \in M$, $n \in N$; consequently,

$$\iint_{MN} f(m,n) + f(n,m)\,dm\,dn \;>\; \iint_{MN} g(m,n) + g(n,m)\,dm\,dn. \tag{24}$$

Moreover, since—by assumption—$f(p,p) = g(p,p)$ and $f(q,q) = g(q,q)$, we can tailor the neighborhoods M and N to make the difference

$$\left[\iint_{MM} f(m,m')\,dm\,dm' + \iint_{NN} f(n,n')\,dn\,dn'\right] - \left[\iint_{MM} g(m,m')\,dm\,dm' + \iint_{NN} g(n,n')\,dn\,dn'\right] \tag{25}$$

as small as we want. Consider, then, the function $h:\mathbf{R} \to \mathbf{R}$ defined by

$$h(u) = \begin{cases} 1 & \text{if } u \in M \cup N, \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

Again, $h \bullet$ is a pointwise transformation. However, $h \bullet I$ fails again to be microbalanced because, for suitably tailored M and N,

$$\begin{aligned} E[h(\kappa)\,h(\lambda)] &= \iint_{MM} f(u,v)\,du\,dv + \iint_{NN} f(u,v)\,du\,dv + \iint_{MN} f(u,v) + f(v,u)\,du\,dv \\ &> \iint_{MM} g(u,v)\,du\,dv + \iint_{NN} g(u,v)\,du\,dv + \iint_{MN} g(u,v) + g(v,u)\,du\,dv \;=\; E[h(\gamma)\,h(\nu)]. \end{aligned} \tag{27}$$ ∎

3.3. Corollary. *Let I be a random stimulus such that for all $(x,y,t),(x',y',t') \in \mathbf{Z}^3$, the pair $(I[x,y,t], I[x',y',t'])$ has a continuous joint density. Then I is microbalanced under all pointwise transformations if the following condition holds for all $x,y,t,x',y',t' \in \mathbf{Z}$. For f the joint density*

of $(I[x,y,t], I[x',y',t'])$, *and* g *the joint density of* $(I[x,y,t'], I[x',y',t])$, *either*

$$f(p,q) = g(p,q) \qquad \text{for all } p,q \in R, p \neq 0, q \neq 0, \tag{28}$$

or

$$f(p,q) = g(q,p) \qquad \text{for all } p,q \in R, p \neq 0, q \neq 0. \tag{29}$$

Proof. If Eq. (28) holds for some $(x,y,t),(x',y',t') \in Z^3$, then we also have

$$f(q,p) = g(q,p) \qquad \text{for all } p,q \in R, p \neq 0, q \neq 0, \tag{30}$$

and we obtain Eq. (19) by adding Eq. (28) and Eq. (30). The same reasoning applies for Eq. (29). ∎

A random stimulus microbalanced under all pointwise transformations, but quite different from J of example 3.1 is the following, suggested by J. Lappin (1989).

3.4. Stimulus K: Rotating random dot cylinder. Construct K by taking the parallel projection of a set of points on (an/or inside) the surface of a cylinder rotating around a vertical axis. Let the contrast values of the points be independent, identically distributed random variables. As is well known, when properly constructed, K can display a very strong kinetic depth effect, with dots moving in one direction seen as being in the front of the axis of rotation, and dots moving in the other direction seen as being in the back (Ullman, 1979; Dosher, Landy, & Sperling, 1989). Nonetheless, K is microbalanced under all pointwise transformations: All of K's systematic motion is horizontal; thus, we can drop reference to y, and note that for any x,t,x',t', the joint distribution of $(K[x,t], K[x',t'])$ is identical to that of $(K[x,t'], K[x',t])$. Hence, by Corollary 3.3, Condition (3), K is microbalanced under all pointwise transformations.

4. Texture quilts.

The rest of this paper is devoted to illustrating how the results of Section 3 can be applied to construct stimuli which display consistent apparent motion that cannot be exposed to standard

analysis by any purely temporal transformation. Specifically, we shall demonstrate several motion-displaying stimuli, called *texture quilts* (Definition 4.1), that are microbalanced under all purely temporal transformations.

FIG 3

As illustrated in Fig. 3, the simplest transformations that suffice to expose the motion of texture quilts to standard analysis involve a purely spatial linear filter $s\bullet$ followed by a rectifier $r\bullet$:

$$T(Q) = r\bullet(s \bullet Q). \tag{31}$$

The spatial filter $s\bullet$ will respond with varying energy throughout regions of the visual field, depending on whether or not the textures to which it is tuned populate those regions. However, the output of a linear filter to a texture is positive or negative depending on the local phase of the texture. The purpose of rectification is to transform regions of high-variance $s\bullet$ response into regions of high average value, thus insuring that the rectified output registers the presence or absence of texture, independent of phase. The result $T(Q)$ is a spatiotemporal function whose value reflects the local texture preferences of $s\bullet$ in the visual field as a function of time (Bergen & Adelson, 1988; Caelli, 1985).[2]

The essential trick in all the quilt examples we consider is to patch together various brief displays of static, random texture, taking appropriate measures to ensure that the resultant stimulus satisfies the following definition.

4.1. Definition of a texture quilt. Let $A \subset \mathbb{Z}^2$ be a set of points in space, and let $t_0, t_1, \ldots, t_N$ be a strictly increasing sequence of times, with $T = \{t \mid t_0 \le t < t_N\}$. Call any random stimulus Q satisfying the following conditions a *texture quilt*:

(i) Q assigns 0 to all points outside $A \times T$.

[2] In general, a spatial linear filter followed by a pointwise nonlinearity can have arbitrarily high order Volterra kernels, depending on the order of the Taylor series of the pointwise transformation. However, if we take the rectifier of step (2) to be $Rect(x) = x^2$, then the squared output of a spatial filter is a second order spatial transformation. Standard motion analysis is yet another second order transformation. Thus, when we subject the squared filter output to standard motion analysis, we are applying a fourth order operator.

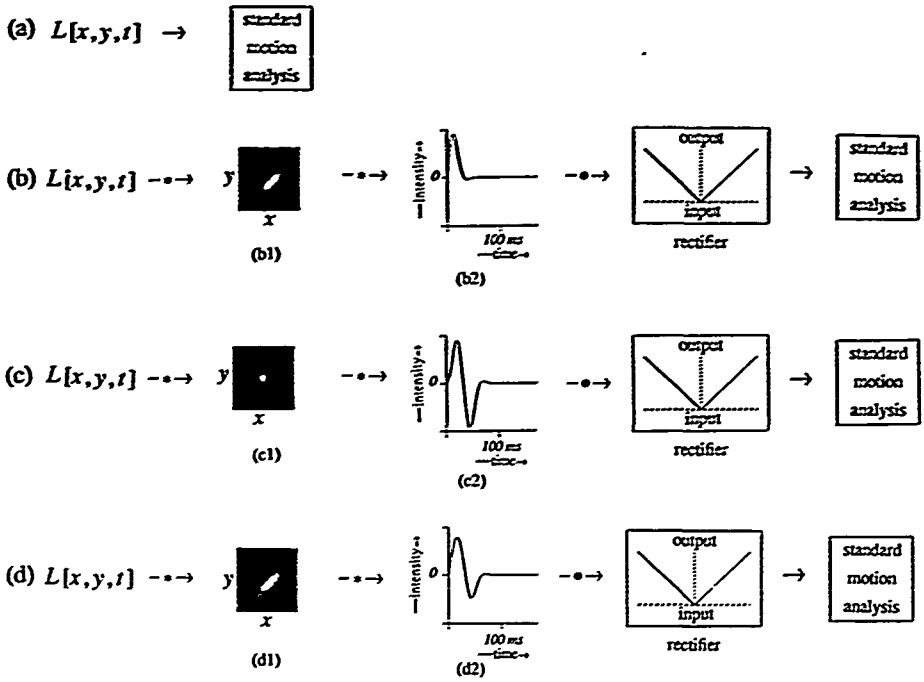


Fig. 3. Fourier and nonFourier motion mechanisms. (a) Fourier motion mechanisms apply standard motion-analysis directly to the luminance signal L. (b, c, d) NonFourier mechanisms apply standard motion analysis to a nonlinear transformation of luminance. (b) A simple nonFourier mechanism applies a signal transformation comprised of a spatiotemporal linear filter, followed by a pointwise nonlinearity. The * 's indicate spatial and temporal convolution, respectively, and • indicates multiplication. The filtering performed in (b) is roughly pointwise in time (the temporal impulse response b2 approximates an impulse), and the nonlinearity applied is a full-wave rectifier. This system (with appropriately chosen spatial filter, b1) will extract the motion of the texture quilts shown in Figs. 4b, 5d, 6c, and 6d. It will not extract the motion of stimulus J, the traveling contrast-reversal of the random vertical bar pattern shown in Fig. 2a. (c) A spatially pointwise (the spatial impulse response c1 approximates an impulse), system with a flicker-sensitive temporal filter and a full-wave rectifier. Because of the flicker sensitivity, this mechanism will extract the motion of the traveling contrast-reversal of the random vertical bar pattern shown in Fig. 2a but not the motion of the texture quilts shown in Figs. 4b, 5d, 6c, and 6d. (d) The temporal filter d2 averages the temporal filters b2 and c2, and the pointwise nonlinearity is a full-wave rectifier. With an appropriate spatial filter d1, this nonFourier system extracts the motion of any corresponding texture quilt as well as the motion of the traveling contrast-reversal of the random vertical bar pattern shown in Fig. 2a. However, it would be less well-suited to these tasks than the detectors shown in (b) and (c) whose temporal filters it averages.

(ii) For $i = 0, 1, \ldots, N-1$, the random values assigned by Q to points in A at time t_i remain unchanged until time t_{i+1}.

(iii) *Independence.* For $i = 0, 1, \ldots, N-1$, the random substimuli Q^i, defined, for all points α in space and all times t, by

$$Q^i[\alpha, t] = \begin{cases} Q[\alpha, t] & t_i \le t < t_{i+1},\ \alpha \in A \\ 0 & \textit{otherwise} \end{cases} \tag{32}$$

are jointly independent.

(iv) *Symmetry.* For any $\alpha, \beta \in A$, and any $t \in T$, the joint distribution of $(Q[\alpha, t], Q[\beta, t])$ is identical to the joint distribution of $(Q[\beta, t], Q[\alpha, t])$.

Terminology: Call A and T respectively Q's *spatial* and *temporal regions of activity*, and for $i = 0, 1, \ldots, N-1$, call $\{t \mid t_i \le t < t_{i+1}\}$ the i^{th} *timeblock* of Q.

The empirical usefulness of texture quilts derives from proposition 4.3 in conjunction with the fact that it is easy to construct various sorts of texture quilts which display consistent apparent motion across independent realizations. The proof of proposition 4.3 is eased by the following

4.2. Lemma. *Let Q be a texture quilt with spatial region of activity A. Then for any $\alpha, \beta \in A$, the pair of temporal functions (Q_α, Q_β) is distributed identically to the reverse pair (Q_β, Q_α).*

Proof. From Definition 4.1 (i) and (ii), note that for temporal functions P and R, the density of the joint assignment $(Q_\alpha, Q_\beta) = (P, R)$ is 0 unless each of P and R is constant throughout each timeblock, and 0 outside T. Thus, any P and R for which the joint assignment $(Q_\alpha, Q_\beta) = (P, R)$ has nonzero density are completely determined by the values $P[t_i] = p_i$, and $R[t_i] = r_i$, for $i = 0, 1, \ldots, N-1$; For f_i the joint density of $(Q_\alpha[t_i], Q_\beta[t_i])$, Definition 4.1 (iii) thus implies that the density of the joint assignment $(Q_\alpha, Q_\beta) = (P, R)$ is

$$\prod_{i=0}^{N-1} f_i(p_i, r_i). \tag{33}$$

But by Definition 4.1 (iv), the quantity (33) is equal to

$$\prod_{i=0}^{N-1} f_i(r_i, p_i), \tag{34}$$

which is the density of the reverse occurrence that $(Q_\beta, Q_\alpha) = (P, R)$. ∎

4.3. Texture quilts are microbalanced under purely temporal transformations.

I. Any texture quilt with a continuous joint density is microbalanced under all purely temporal, continuous transformations.

II. Any discretely distributed texture quilt is microbalanced under all purely temporal transformations.

Proof of I. Let Q be a texture quilt with a continuous joint density, and let Φ be an arbitrary purely temporal, continuous transformation. We must prove that $\Phi(Q)$ is microbalanced. We can, of course, accomplish this by proving that $\Phi(Q)$ is microbalanced under all pointwise transformations (since, in particular, the identity transformation is pointwise). This turns out to be a convenient approach.

Let α, β be points in space, and let t and u be points in time. Because Φ is bounded and continuous and Q has a continuous joint density, we know that the joint density f of $(\Phi(Q)[\alpha, t], \Phi(Q)[\beta, u])$ and the joint density g of $(\Phi(Q)[\beta, t], \Phi(Q)[\alpha, u])$ both exist and are continuous on $\mathbf{R}^2$. We shall show for any $(p, r) \in \mathbf{R}^2$ with neither p nor r equal to 0, that either $f(p, r) = g(p, r)$ or $f(p, r) = g(r, p)$. The proposition will then follow from corollary 3.3.

Case 1: At least one of α or β is outside A. Suppose α is outside A. Then by Definition 4.1 (i), $Q_\alpha = 0$; hence $\Phi(Q)[\alpha, t] = \Phi(Q)[\alpha, u] = 0$. Consequently, $f(p, r) = g(r, p) = 0$ whenever $p \neq 0$. Thus Eq. (29) holds vacuously, with

$$f(p, r) = g(r, p) = 0 \qquad \text{for all } p, r \in \mathbf{R}, p \neq 0, r \neq 0. \tag{35}$$

Case 2: Both α and β are in A. Let F be the joint density of (Q_α, Q_β) and G the joint density

of (Q_β, Q_α). By lemma 4.2, $F = G$. Clearly, then, for F_Φ the joint density of $(\Phi(Q_\alpha), \Phi(Q_\beta))$ and G_Φ the joint density of $(\Phi(Q_\beta), \Phi(Q_\alpha))$, it follows that $F_\Phi = G_\Phi$. For any $p, r \in \mathbf{R}$, recall that $f(p,r)$ is the density of the co-occurrence that $\Phi(Q)[\alpha, t] = p$, and $\Phi(Q)[\beta, u] = r$, but this is precisely the density of the event that $(\Phi(Q_\alpha)[t], \Phi(Q_\beta)[u]) = (p,r)$. This density, however, is equal to the integral of F_Φ over all pairs of temporal functions (P,R) such that $P[t]=p$ and $R[u]=r$. Similarly, $g(p,r)$ is the density of the co-occurrence that $\Phi(Q)[\beta, t] = p$, and $\Phi(Q)[\alpha, u] = r$, but this is the density of the event that $(\Phi(Q_\beta)[t], \Phi(Q_\alpha)[u]) = (p,r)$, which is equal to the integral of G_Φ over all pairs of temporal functions (P,R) such that $P[t]=p$ and $R[u]=r$. However, as we have already noted, $F_\Phi = G_\Phi$, implying that $f = g$. Apply corollary 3.3 to complete the proof. ∎

The proof of II is similar.

The rest of Section 4 is devoted to showing how to construct two kinds of simple texture quilts. In Section 5, we apply these construction techniques in an experiment to investigate what sorts of textural characteristics are actually processed for motion information by the visual system.

4.4. Binary texture quilts.

4.4.1. A general technique for constructing binary texture quilts. The simplest sorts of texture quilts involve only two contrast values. As in Definition 4.1, let $T = \{t \parallel t_0 \le t < t_N\}$ be the temporal region of activity, with new timeblocks beginning at times $t_0, t_1, \ldots, t_{N-1}$. Let A be the spatial region of activity. Associate with timeblocks $i = 0, 1, \ldots, N-1$ spatial functions f_i (called *timeblock pictures*), each of which is 0 everywhere outside A, and takes only the values 1 and −1 within A. In addition, associate with timeblocks *0* through *N-1* a family

$$\phi_0, \phi_1, \ldots, \phi_{N-1} \tag{36}$$

of jointly independent random variables, each of which takes the value 1 or −1 with equal probability

Then, for $i = 0, 1, \ldots, N-1$, set

$$B_i[x, y, t] = \begin{cases} f_i[x, y] & \text{if } t \text{ is in timeblock } i, \\ 0 & \text{otherwise,} \end{cases} \tag{37}$$

and construct the random stimulus

$$B = \phi_0 B_0 + \phi_1 B_1 + \ldots + \phi_{N-1} B_{N-1}. \tag{38}$$

It is easy to see that B is a texture quilt. First, the functions B_i are defined to satisfy Definition 4.1 (i) and (ii). The joint independence of the random variables ϕ_i ensures that B satisfies Definition 4.1 (iii). To see that Definition 4.1 (iv) is satisfied, note that for any $\alpha, \beta \in A$, either (i) $B_i[\alpha, t_i] = B_i[\beta, t_i]$ or (ii) $B_i[\alpha, t_i] = -B_i[\beta, t_i]$. In case (i),

$$B[\alpha, t_i] = \phi_i B_i[\alpha, t_i] = \phi_i B_i[\beta, t_i] = B[\beta, t_i], \tag{39}$$

implying that the pair $(B[\alpha, t_i], B[\beta, t_i])$ is distributed identically to the pair $(B[\beta, t_i], B[\alpha, t_i])$ (each pair with an equal probability of taking the value (1, 1) or (−1, −1)). In case (ii)

$$B[\alpha, t_i] = -B[\beta, t_i], \tag{40}$$

and the pair $(B[\alpha, t_i], B[\beta, t_i])$ is distributed identically to the pair $(B[\beta, t_i], B[\alpha, t_i])$, each with an equal probability of assuming the value (1, −1) or (−1, 1). Thus Definition 4.1 (iv) is satisfied along with 4.1 (i), (ii) and (iii).

4.4.2. Stimulus: The sidestepping, randomly contrast-reversing, vertical edge. In Fig. 4b are displayed the 9 timeblock pictures comprising a particularly simple binary texture quilt. Note that the vertical dimension of Fig. 4b combines time and vertical space, precisely as a strip of movie film, scanned vertically, combines time & space. Timeblock pictures are separated by grey lines. Fig. 4a shows the timeblock pictures f_0 through f_8 used in the construction. f_0 assigns the value −1 to all points (x, y) of the horizontal rectangle comprising the spatial region of activity, A. f_1 assigns 1 to the points in the leftmost eighth of A, and −1 to the points in the right seven eighths. The timeblock pictures f_2 through f_8 continue to shift the vertical edge rightward through A until, in picture 8, A is uniformly 1. Multiplying each timeblock picture $i = 1, 2, \ldots, 9$ by its associated random variable ϕ_i

yields, in this particular realization, the stimulus given in Fig. 4b.

FIG 4

The construction of the side-stepping contrast-reversing edge (Fig. 4b) is symmetric to the construction of the traveling contrast-reversal of a random black-or-white vertical bar pattern (J in Fig. 2a). Transposing the x and t dimensions in Fig. 4b gives the xt-cross section of a random stimulus J (e.g., Fig. 2a). This stimulus exhibits an unusual symmetry between space and time. Whereas the texture quilt of Fig. 4b is microbalanced under all purely temporal transformations, its transpose J (Fig. 2b) is microbalanced under all *purely spatial* transformations. Extracting motion from J requires *temporal* filtering followed by a nonlinearity. This process is essentially different from the process by which motion is extracted from texture quilts (e.g., Figs. 4b, 7a, 7b and 7c) which requires a *spatial* nonlinearity.

4.4.3. Stimulus: Oppositely oriented static squarewaves selected by a drifting grating. Figure 5d shows the four timeblock pictures comprising another binary texture quilt constructed using technique 4.4.1. In Fig. 5a is shown a probabilistically defined sinewave grating, a stimulus whose motion is readily extracted by standard motion analysis. In Figs. 5b1 and 5b2 are shown static vertical and horizontal squarewave gratings. The stimulus of Fig. 5c is obtained by using Fig. 5a to select between the vertical and horizontal gratings of Figs. 5b1 and 5b2. If the function of Fig. 5a is 1 at a certain point in space-time, the corresponding point in Fig. 5c is assigned the value of the corresponding point in Fig. 5b1; otherwise the point in Fig. 5c is assigned the value of the corresponding point in Fig. 5b2. Although Figs. 5c and 5d look similar, they differ in an important respect: the stimulus of Fig. 5d is microbalanced under all purely temporal transformations, while that of Fig. 5c is not microbalanced It is possible to design Fourier mechanisms to detect the motion of Fig. 5c, but not that of Fig. 5d. The critical difference is that the timeblock pictures of Fig. 5d are jointly independent, while those of Fig. 5c are not: Fig. 5d is obtained by randomly reversing the contrasts of the timeblock pic-

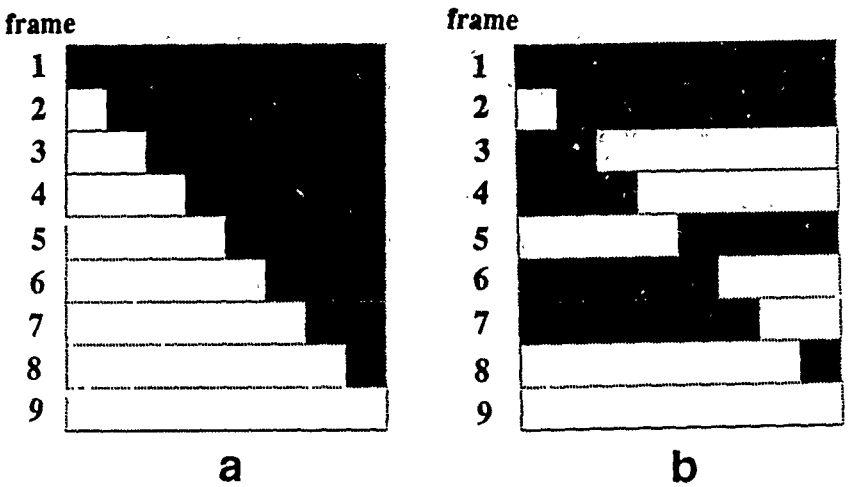


Fig. 4. Edge-driven motion from an ordinary edge and from a binary texture quilt. (a) A rightward moving light-dark edge visible to Fourier and nonFourier motion systems. Nine entire frames are shown; each frame consists of an area of contrast +1 and area of contrast -1. (b) A realization of the *sidestepping, randomly contrast-reversing vertical edge*. This random stimulus is a texture quilt and hence microbalanced under all purely temporal transformations: that is, its rightward motion would be inaccessible to standard motion analysis even if this analysis were preceded by an arbitrary, purely temporal transformation. Each frame of (b) was derived from the corresponding frame of (a) by multiplying the entire frame by a random variable that takes the value 1 or −1 with equal probability. The frame random variables are jointly independent. A straightforward way to extract the motion of this texture quilt is to (i) apply a linear filter sensitive to vertical edges, (ii) rectify the filtered output, and (iii) submit the result to standard motion analysis.

tures of Fig. 5c.

FIG 5

4.5. Sinusoidal texture quilts.

It is not difficult to elaborate technique 4.4.1 to a method for constructing quilts involving textures of arbitrarily many contrast values. We illustrate the principle in the construction of quilts comprised of patches of sinusoidal grating.

4.5.1. A general technique for constructing sinusoidal texture quilts. As in Definition 4.1, let $T = \{t \mid t_0 \le t < t_N\}$ be the temporal region of activity, with new timeblocks beginning at times $t_0, t_1, \ldots, t_{N-1}$. Let A be the spatial region of activity. Associate with timeblocks $i = 0, 1, \ldots, N-1$, spatial functions W_i, each of which is 0 everywhere outside A, and takes only the values 1 and -1 within A. The stimulus in each time block will be composed of two components characterized by spatial frequencies (ω_i, θ_i) and $(\tilde{\omega}_i, \tilde{\theta}_i)$, respectively, and independent phases ρ_i, $\tilde{\rho}_i$, respectively. Let

$$\omega_0, \theta_0, \tilde{\omega}_0, \tilde{\theta}_0, \omega_1, \theta_1, \tilde{\omega}_1, \tilde{\theta}_1, \ldots, \omega_{N-1}, \theta_{N-1}, \tilde{\omega}_{N-1}, \tilde{\theta}_{N-1} \tag{41}$$

be integers. Let P be an integer, and let

$$\rho_0, \tilde{\rho}_0, \rho_1, \tilde{\rho}_1, \ldots, \rho_{N-1}, \tilde{\rho}_{N-1} \tag{42}$$

be jointly independent random variables, each uniformly distributed on the set $\{0, 1, \ldots, P-1\}$. Then, define the stimulus S as the sum of N component stimuli S_i defined in each timeblock:

$$S = \sum_{i=0}^{N-1} S_i, \tag{43}$$

where, for $i = 0, 1, \ldots, N-1$, S_i is zero everywhere outside timeblock i; and for all t in timeblock i,

$$S_i[x, y, t] = f_i[x, y] = \begin{cases} \cos(2\pi(\omega_i x + \theta_i y - \rho_i)/P) & \text{if } W_i[x, y] = 1, \\ \cos(2\pi(\tilde{\omega}_i x + \tilde{\theta}_i y - \tilde{\rho}_i)/P) & \text{if } W_i[x, y] = -1, \\ 0 & \textit{otherwise.} \end{cases} \tag{44}$$

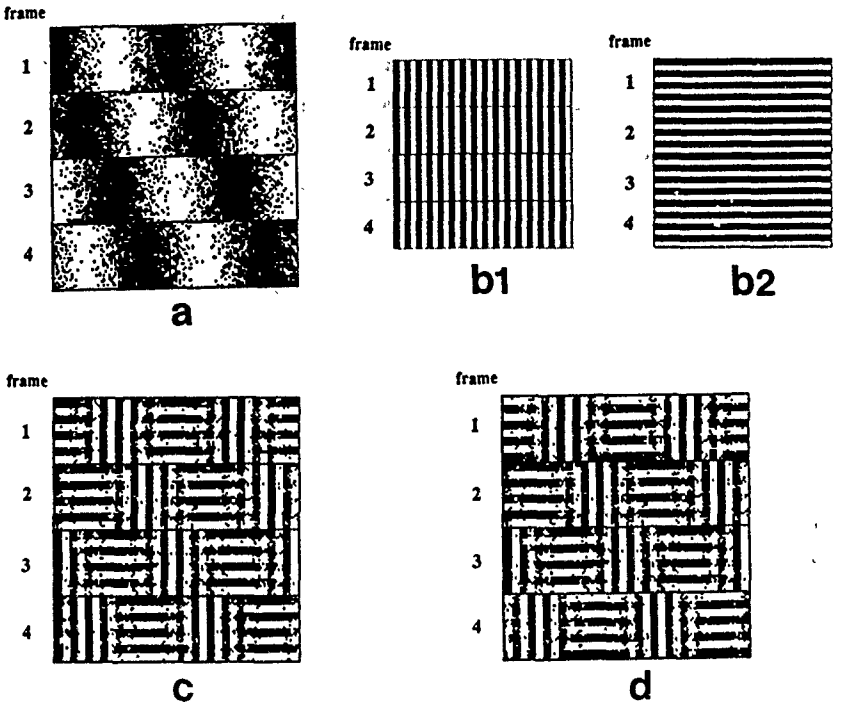


Fig. 5. Orientation-driven nonFourier motion from a binary texture quilt. (a) A probabilistically defined sinewave grating that steps rightward 90 degrees between frames. The rightward motion in (a) is accessible to all motion detectors. (b1) Four frames of a static, vertical squarewave grating. (b2) Four frames of a static horizontal squarewave grating. (c) A rightward translating texture pattern. For every white point in (a), the corresponding value in (c) is chosen from the vertical square-wave grating in (b1); for every black point in (a), the corresponding value in (c) is chosen from the horizontal square-wave grating in (b2). (c) is not microbalanced; standard motion-analyzers can be designed to detect its motion. (d) A texture quilt. The frames of (d) are derived by multiplying the corresponding frames of (c) by jointly independent random variables, each of which takes the value 1 or −1 with equal probability. The texture quilt (d) is microbalanced under all purely temporal transformations, and therefore its rightward motion is unavailable to any mechanism that applies standard motion analysis to a purely temporal transformation of the visual signal.

It is easy to check that S satisfies Definition 4.1 (i) and (ii). The joint independence of the random phase variables ρ_i, $\tilde{\rho}_i$, for $i = 0, 1, ..., N-1$ entails Definition 4.1 (iii).

It remains to check that S satisfies Definition 4.1 (iv). Consider points $\alpha, \beta \in A$. If $W_i[\alpha] \neq W_i[\beta]$, then, as is easily checked, $S[\alpha, t_i]$ and $S[\beta, t_i]$ are independent and identically distributed (each assuming a value from among $\{\cos(2\pi p/P) \mid p = 0, 1, ..., P-1\}$ with equal probability). On the other hand, if $W_i[\alpha] = W_i[\beta]$, then the pair $(S[\alpha, t_i], S[\beta, t_i])$ is distributed identically to the pair $(S[\beta, t_i], S[\alpha, t_i])$ as a consequence of the following

Lemma. *Let $P \in \mathbb{Z}$, and let $\alpha = (\alpha_x, \alpha_y)$, $\beta = (\beta_x, \beta_y)$ and $\omega = (\omega_x, \omega_y)$ all be elements of $\mathbb{Z}^2$. Then for any integer $p \in \{0,1,...,P-1\}$, there exists an integer $q \in \{0, 1, ..., P-1\}$ such that (writing $\cdot$ for dot product)*

$$\cos(2\pi(\omega\cdot\alpha - p)/P) = \cos(2\pi(\omega\cdot\beta - q)/P) \tag{45}$$

and

$$\cos(2\pi(\omega\cdot\beta - p)/P) = \cos(2\pi(\omega\cdot\alpha - q)/P). \tag{46}$$

Proof. As the reader may check, this is true for $q = (\omega\cdot\alpha + \omega\cdot\beta - p)$ *modulo* P. ∎

Thus, for α, β such that $W_i[\alpha] = W_i[\beta]$, we observe that for any outcome $\rho_i = p$, there exists an equally likely outcome $\rho_i = q$, such that

$$\left\{\cos(2\pi(\omega_i\cdot\alpha - p)/P), \cos(2\pi(\omega_i\cdot\beta - p)/P)\right\} = \left\{\cos(2\pi(\omega_i\cdot\beta - q)/P), \cos(2\pi(\omega_i\cdot\alpha - q)/P)\right\} \tag{47}$$

We infer that the pair $(S[\alpha, t_i], S[\beta, t_i])$ is distributed identically to the pair $(S[\beta, t_i], S[\alpha, t_i])$.

4.5.2. Stimulus: Oppositely oriented static sinusoids selected by a drifting grating. The sinusoidal analog to the binary texture quilt of Fig. 5d is shown in Fig. 6b. In Fig. 6a are shown the functions W_1, W_2, W_3, and W_4 used to select between horizontal and vertical gratings. For this quilt, $\tilde{\omega}_i = \theta_i = 0$, for $i = 1, 2, 3, 4$; and for some integer F (with F/P the number of cycles per pixel), $\omega_i = \tilde{\theta}_i = F$. The texture quilt of Fig. 6b modulates textural orientation across space and time. Alternatively, we can just as easily keep orientation constant and vary spatial frequency.

FIG 6

4.5.3. Stimulus: Static sinusoids of different spatial frequencies, selected by a drifting grating. Figure 6c shows a texture quilt using the sampling functions of Fig. 6a, but setting $\omega_i = \theta_i = 2\tilde{\omega}_i = 2\tilde{\theta}_i$ for $i = 1, 2, ..., 4$.

5. What aspects of texture does the visual system process for motion?

In this section, we describe a psychophysical experiment investigating the question of what characteristics of spatial texture are analyzed for motion information by the visual system. Three texture quilts are compared across four different viewing conditions. These conditions comprise a sequence of similar, but increasingly challenging motion discrimination tasks.

5.1. Procedure. Every texture quilt used in this experiment is comprised of a sequence of jointly independent timeblocks, each lasting 1/30 sec. (Each timeblock consists of two identical refreshes at 1/60 sec.) Each texture quilt is stochastically periodic with a period of 8 timeblocks: that is, for any integer i, the i^{th} timeblock is identically distributed to the $i + 8^{th}$ timeblock. Accordingly, we refer to eight timeblocks of the texture quilt as one *cycle*. The motion elicited by each quilt is carried by a squarewave that selects between two textures, and steps 1/4 cycle on every odd timeblock. The squarewave thus completes one of its four-step cycles in each 8 timeblock cycle of the quilt.

On each trial, a texture quilt moving randomly left or right is presented, and the subject is required to signal (with a button-press) which way the quilt appeared to move. The subject is asked to maintain fixation on a small spot present in the middle of the stimulus throughout the display, and receives feedback after each trial. For each quilt under each viewing condition, the subject performs 100 practice trials followed directly by 100 actual trials. Quilt realizations are jointly independent across trials. The starting phase of the quilt is chosen randomly on each trial.

The four viewing conditions. For a given quilt, the four viewing conditions differ with respect

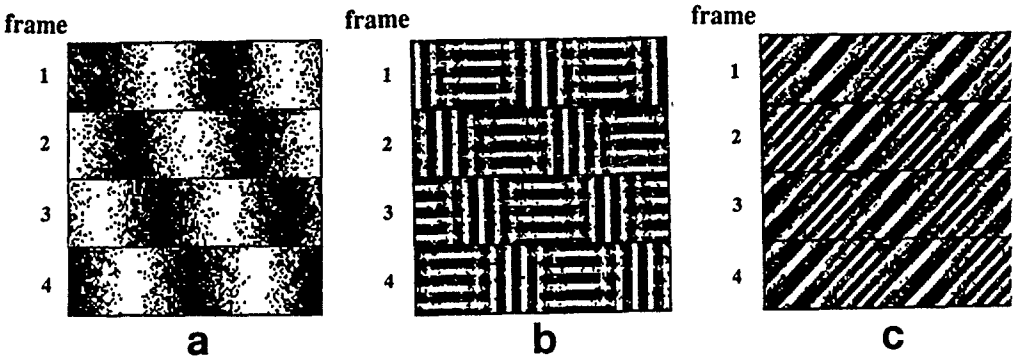


Fig 6 Sinusoidal texture quilts: Motion driven by differences in *orientation* and in *spatial frequency*. (b) and (c) show realizations of random stimuli, each of which is microbalanced under all purely temporal transformations. Their rightward motion cannot be detected by any mechanism that applies standard motion analysis to a purely temporal transformation of the signal. In each case, the 4 frames in (a) select between two sinusoidal patterns. The phases of sinusoids are jointly independent across frames and across different-frequency sinusoidal components patched together in the same frame. The sinusoids mixed in (b) differ in orientation, whereas the sinusoids mixed in (c) have the same orientation, but differ in spatial frequency.

to the number of quilt cycles displayed. In Condition 1, the easiest condition, the subject sees two quilt cycles (each cycle comprised of eight stimulus timeblocks), with each timeblock displayed for 1/30 sec. In Conditions 2, 3, and 4, the subject sees 1.5, 1, and .5 quilt cycles, respectively.

5.1.1. Three quilt stimuli. The first quilt (the F-quilt) modulates textural spatial frequency as a function of space and time, while keeping orientation constant. The eight timeblocks comprising one full cycle of the F-quilt are shown in Fig. 7a. A second quilt (the O-quilt, Fig. 7b) modulates textural orientation as a function of space and time, while keeping spatial frequency constant. A third quilt (the E-quilt, Fig. 7c) spatiotemporally modulates texture between jointly independent binary noise and the so-called "even" texture (Julesz, Gilbert & Victor, 1978).

All stimuli were viewed from 1 m against a mean luminant background. At this distance, each quilt spanned 6.8 horizontal and 3.2 vertical degrees, and the modulating square wave moved at an average velocity of 12.75 deg/sec.

FIG 7

5.1.2. Why these three quilts. In each of the three quilts, a squarewave with vertical bars is used to modulate between two textures as a function of space and time. The squarewave has a spatial frequency of.3 c/deg., and steps 1/4 cycle rightward on every odd timeblock (temporal frequency 3.75 *Hz*, velocity 12.75 deg/sec). We use a 1/4-cycle stepping squarewave to modulate between the two textures comprising each quilt in order to rule out the possibility that the motion elicited by the quilt is being carried by the border between textural regions. That is, the 1/4-cycle stepping squarewave has the advantage that the signal derived from the borders between texture regions is ambiguous in motion content. Given the requirement of 1/4 cycle steps, we changed the particular instantiation of the quilt on even timeblocks (i e., within steps of the squarewave) in order to spread textural energy broadly in temporal frequency without altering the spatial frequency content of the texture.

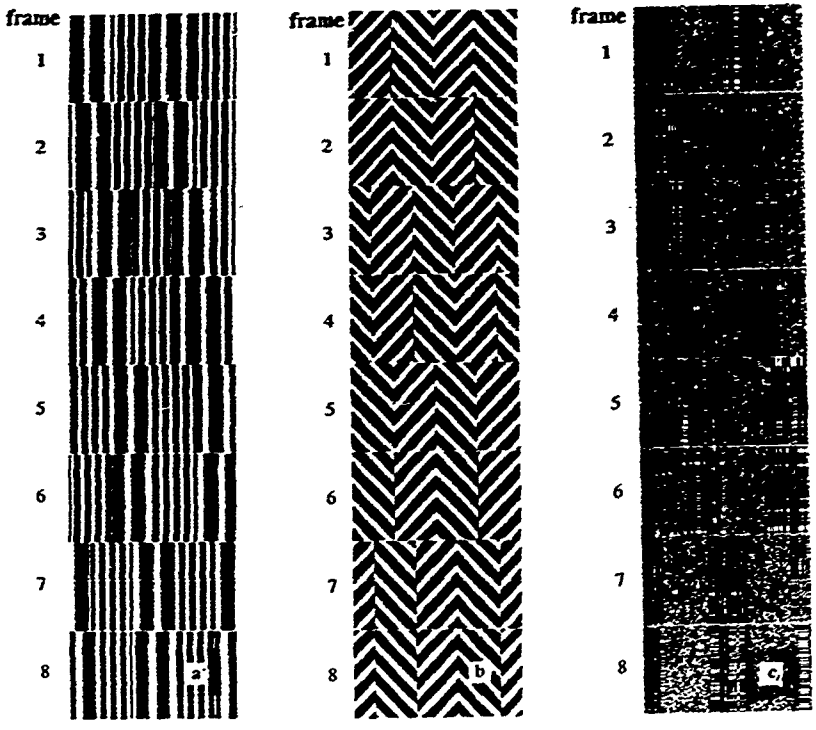


Fig. 7. Three quilts used to study motion carried by modulation of texture spatial frequency, by texture orientation, and by higher order textural characteristics. (a) Eight frames that comprise one cycle of the F-quilt. Motion is generated by by a squarewave modulation of textural spatial frequency. The squarewave grating selects between vertical sinusoidal gratings of spatial frequency 1.2 c/deg and 2.4 c/deg. The texture-modulating squarewave is 0.3 c/deg, and steps 1/4 cycle rightward on every odd frame. Every even frame is independent of and distributed identically to the preceding frame. Presentation proceeds at the rate of 30 frames/sec. This gives the texture-modulating squarewave a temporal frequency of 3.75 Hz and a mean velocity of 25 deg/sec.

(b) Eight frames that comprise once cycle of the O-quilt. In the O-quilt, textural orientation is modulated by the same squarewave used to modulated spatial frequency in the F-quilt. The O-quilt squarewave selects between oppositely oriented sinusoidal gratings that have a spatial frequency of 2.8 c/deg.

(c) Eight frames that comprise once cycle of the E-quilt. In the E-quilt, the texture-modulating squarewave selects between jointly independent binary noise and an "even" texture (Julesz, Gilbert & Victor, 1978) Despite the evident difference between these two textures, every time-independent linear filter has the same expected power for both textures Thus, if motion-from-texture resulted from applying a simple squaring transformation to the output of a spatial linear filter and submitting the result to standard motion analysis, the motion of the E-quilt would be invisible.

It has been previously observed (Watson & Ahumada, 1983a; Ramachandran, Ginsburg & Anstis, 1983; Green, 1986) that motion is carried more effectively by spatiotemporal variation of textural spatial frequency than by variation of textural orientation. The F-quilt and O-quilt were chosen to further investigate this claim. The E-quilt is of interest because the two textures of which it is composed (jointly independent binary noise and the even texture) have identical second order statistics. That is, the joint distribution of any given pair of points in space is the same under both the component textures of the E-quilt. This means that, despite the obvious difference in appearance between the component textures, the expected energy in the response of any given spatial linear filter is the same for both component textures. If the pointwise nonlinearity applied to the output of the spatial linear filter prior to motion analysis were simple squaring, it would be impossible to detect the motion of the E-quilt.

Victor and Conte (1990) studied apparent motion elicited by E-quilts, and noted that it is much weaker than motion elicited by comparable stimuli (also texture quilts) that modulate between textures differing in spatial frequency. Our experiment confirms this finding.

5.2. Results Two subjects participated in the study, *CC* (the experimenter) and *GA* (naive). The results for *CC* are shown in Fig. 8b and those for *GA* are shown in Fig. 8a. Note first that both subjects were able to reliably discriminate left/right motion in all three stimuli although subject *GA* failed with the E-quilt at the briefest exposure. The two subjects performed comparably well at motion direction discrimination of the O-quilt, but *CC* was much better than *GA* at detecting the motion of both the F-quilt and the E-quilt. Subject *CC* was better at detecting the motion of the F-quilt than the O-quilt; the reverse was true of subject *GA*.

It is possible that these performance differences reflect a genuine differences in the perceptual apparatus of the two subjects. However, we cannot rule out the possibility that the better performance of subject *CC* is due merely to his vastly greater experience with motion perception tasks of this sort.

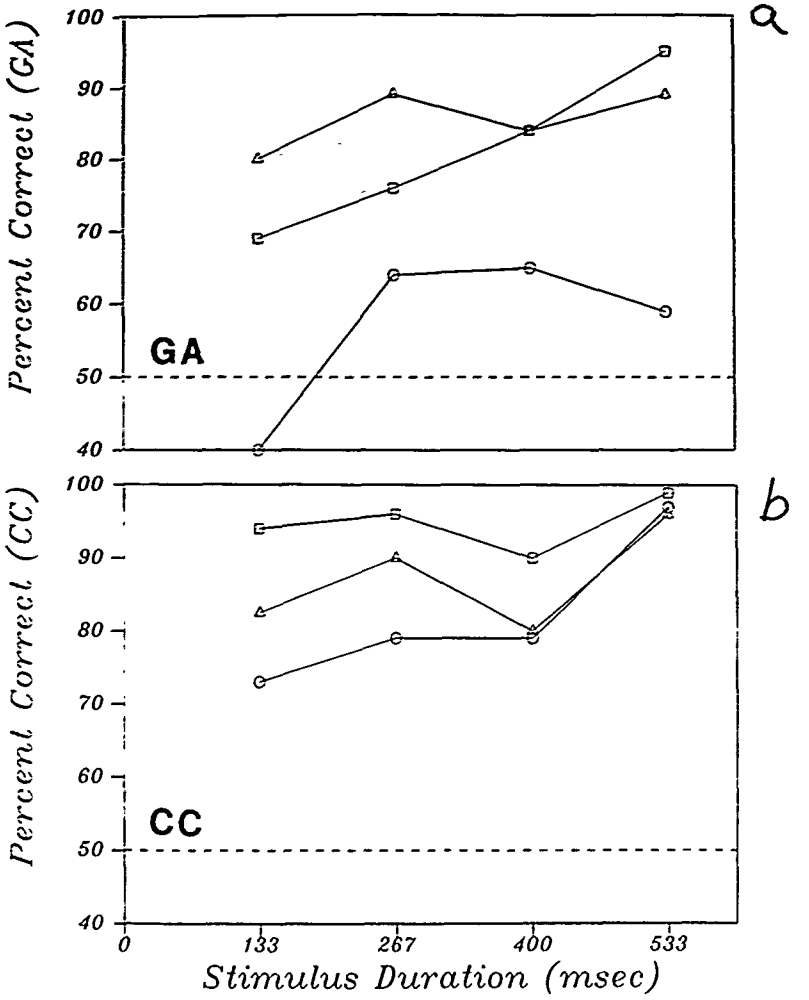


Fig. 8. The percent of correct direction-of-motion judgments to the F-quilt, the O-quilt, and the E-quilt as a function of stimulus duration. The panels show data for subjects CC and GA, respectively Each data point is the mean of 100 judgments. (Squares) F-quilt, (triangles) O-quilt, (circles) E-quilt. The stimulus durations of 133, 266, 400, and 533 ms, correspond to stimulus presentations of 0.5, 1, 1.5 and 2 quilt cycles.

5.3. Discussion. Many of the models proposed to explain rapid, preattentive segregation of spatial textures (Caelli, 1985; Beck, Sutter & Ivry, 1987; Sutter, Beck & Graham, 1989; Bergen & Adelson, 1988; Malik & Perona, 1989) can easily be adapted to deal with the motion displayed by texture quilts. The texture segregation models in this class typically subject the visual input function to a linear transformation (a "texture grabber") followed by a pointwise nonlinearity (such as a rectifier or thresholder) to indicate the presence or absence of the texture. Such models propose that two contiguous textural regions would generate a perceptual boundary if the visual system were equipped with a linear filter that is differentially tuned to one of the textures.

An analogous mechanism to detect the motion of texture quilts, suggested by the current experiment and the work of Victor and Conte (1990), (i) convolves the input stimulus with a spatial texture-grabbing filter tuned to the moving texture, then (ii) squares the output of the filter, to transform regions of high-energy filter output into regions of high average value, and (iii) subjects the rectified output to standard motion analysis. However, the transformation applied in steps (i) and (ii) does not distinguish between the two textures comprising the E-quilt, and therefore fails to account for the good performance with the E-quilt. A simple modification to deal with texture segregation and motion perception of the E-quilt is to assume some other post-filter rectification operation than the squaring operation. It is quite easy to choose a linear filter in combination with a post-filter rectifier (other than the squaring operation) that will segregate the random and even textures (e.g., Julesz & Bergen, 1983). The current experiment does not specifically indicate the kind of rectification that might be involved.

What sorts of filters are available to the visual system to compute motion from texture? For example, Daugman (1985) points out that (i) Gabor filters provide an optimal trade-off between resolution in the space and spatial frequency domains, and (ii) many investigators note that simple cells in cat striate cortex are well-modeled by oriented Gabor filters (e.g., Wilson & Sherman, 1976; DeValois, DeValois & Yund, 1979; Andrews & Pollen, 1979). Are the linear filters that serve motion-from-texture computations Gabor-like cortical simple cells? The theory reported here

provides a tool, and the demonstration experiments illustrate how it might be used to answer such questions.

6. Summary.

The main contributions of this paper are (i) to introduce the notion of a random stimulus *microbalanced under all pointwise transformations*, (ii) to provide necessary and sufficient conditions for a random stimulus to be of this sort, (iii) to use this result to construct apparent motion stimuli called *texture quilts* that are microbalanced under all purely temporal transformations, and (iv) to show that subjects can reliably discriminate the motion direction of three kinds of texture quilts.

Texture quilts provide a flexible array of tools for studying motion perception that is truly mediated by spatiotemporal modulation of spatial texture without contamination by mechanisms responsive to the motion extracted directly by standard analysis or motion extracted by standard analysis of any purely temporal transformation of the stimulus.

Acknowledgements.

The research reported here was supported by USAF Life Science Directorate, Visual Information Processing Program Grants 85-0364 and 88-0140.

References

Adelson, E.H. and J. Bergen (1985) "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Am. A*, 2, 2, 284-299.

Adelson, E.H. and J. Bergen (1986) "The extraction of spatio-temporal energy in human and machine vision," *Proceedings* of the *IEEE Workshop on Motion: Representation and Analysis*, 151-155.

Andrews, B.W. & D.A. Pollen (1979), "Relationship between spatial frequency selectivity and receptive field profile of simple cells," *J. Physiol. (London)*, 287, 163-176.

Anstis, S.M. (1970) "Phi movement as a subtraction process," *Vision Res.*, 10, 1411-1430.

Baker, C.L. & O. Braddick (1982a) "Does segregation of differently moving areas depend on relative

or absolute displacement *Vision Res.*, 22, 851-856.

Baker, C.L. & O. Braddick (1982b) "The basis of area and dot number effects in random dot motion perception," *Vision Res.*, 22, 1253-1260.

Beck, J., A. Sutter & R. Ivry (1987) "Spatial frequency channels and perceptual grouping in texture segregation," *Computer Vision, Graphics, and Image Processing*, 37, 299-325.

Bergen, J.R. and E.H. Adelson (1988) "Early vision and texture perception," *Nature*, 333, *6171*, 363-364.

Bell, H.H. & J.S. Lappin (1979) "The detection of rotation in random dot patterns," *Perception and Psychophysics*, 26, 415-417.

Bowne, S.F., S.P. McKee & D.A. Glaser (1989) "Motion interference in speed discrimination," *J. Opt. Soc. Am. A*, 6, 7, 1112-1121.

Braddick, O. (1973) "The masking of apparent motion in random-dot patterns," *Vision Res* , 13, 355-359.

Braddick, O. (1974) "A short-range process in apparent motion," *Vision Res.*, 14, 519-527.

Caelli, T. (1985) "Three processing characteristics of visual texture segmentation," *Spatial Vision*, 1, *1*, 19-30.

Cavanagh, P. "Motion: The long and the short of it," talk at *Conference on Visual From and Motion Perception: Psychophysics, Computation, and Neural Networks* (Meeting dedicated to the memory of the late Kvetoslav Prazdny), Boston University, Massachusetts, March 5, 1988.

Cavanagh, P., M. Arguin & M. von Grunau (1989), "Interattribute apparent motion," *Vision Res.*, 29, *9*, 1197-1204.

Chang, J.J. & B. Julesz (1983a) "Displacement limits, directional anisotropy and direction versus form discrimination in random dot cinematograms," *Vision Res.*, 23, 639-646.

Chang, J.J. & B. Julesz (1983b) "Displacement limits for spatial frequency random-dot cinematograms in apparent motion," *Vision Res* , 23, 1379-1386.

Chang, J.J. & B. Julesz (1985) "Cooperative and non-cooperative processes of apparent movement of

random dot cinematograms," *Spatial Vision*, 1, *1*, 39-45.

Chubb, C. & G. Sperling (1987) "Drift-balanced random stimuli: A general basis for studying non-Fourier motion perception," *Investigative Ophthalmology and Visual Science*, 28, p. 233.

Chubb, C. & G. Sperling (1988) "Drift-balanced random stimuli: A general basis for studying non-Fourier motion perception," *J. Opt. Soc. Am. A*, 5, *11*, 1986-2007.

Daugman, J.G. (1985) "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Am. A*, 2, 7, 1160-1169.

Derrington, A.M. & D.R. Badcock (1985) "Separate detectors for simple and complex grating patterns?" *Vision Res.*, 25, 1869-1878.

Derrington, A.M. & G.B. Henning (1987) "Errors in direction-of-motion discrimination with complex stimuli," *Vision Res* , 27, 61-75.

DeValois, K.K., R.L. DeValois & E.W. Yund (1979) "Responses of striate cortical cells to grating and checkerboard patterns," *J. Physiol. (London)*, 291, 483-505.

van Doorn, A.J. & J.J. Koenderink (1984) "Spatiotemporal integration in the detection of coherent motion," *Vision Res.*, 24, 47-54.

Dosher, Barbara A., M. S. Landy, & G. Sperling (1989) "Ratings of kinetic depth in multi-dot displays," *Journal of Experimental Psychology: Human Perception and Performance*, 15, 116-425.

Green, M. (1986) "What determines correspondence strength in apparent motion," *Vision Res.*, 26, 599-607.

Julesz, B. (1971) *Foundations of Cyclopean Perception*, Chicago: University of Chicago Press.

Julesz, B., and J.R. Bergen (1983), "Textons, the fundamental elements in preattentive vision and perception of textures," *Bell Systems Technical Journal*, *62:6*, 1619-1645.

Julesz, B , E. Gilbert, and J.D. Victor (1978) "Visual discrimination of textures with identical third-order statistics." *Biological Cybernetics*, 31, 137-140.

Lappin, J.S. (1989), Personal communication, June 20.

Lappin, J.S. & H H. Bell (1972) "Perceptual differentiation of sequential visual patterns," *Perception and Psychophysics*, **12**, 129-134.

Lelkins, A.M.M. & J.J. Koenderink (1984) "Illusory motion in visual displays," *Vision Res.*, **24**, 1083-1090.

Malik, J., and Perona, P. (1989). A computational model of texture perception. Computer Science Division (EECS) Report No. UCB/CSD 89/491, University of California, Berkeley, California.

Nakayama, K. & G. Silverman (1984) "Temporal and spatial characteristics of the upper displacement limit for motion in random dots," *Vision Res.*, **24**, 293-300.

Pantle, A. & K. Turano (1986) "Direct comparisons of apparent motions produced with luminance, contrast-modulated (CM), and texture gratings" *Investigative Ophthalmology and Visual Science*, **27**, *3*, p. 141.

Petersik, J.T., K.I. Hicks & A.J. Pantle (1978) "Apparent movement of successively generated subjective figures," *Perception*, **7**, 371-383.

Ramachandran, V.S. & S.M. Anstis (1983) "Displacement thresholds for coherent apparent motion in random dot-patterns," *Vision Res.*, **23**, 1719-1724.

Ramachandran, V.S., Ginsburg, A. and S.M. Anstis (1983) "Low spatial frequencies dominate apparent motion," *Perception*, **12**, 457-461.

Ramachandran, V.S., V.M. Rao and T.R. Vidyasagar (1973) "Apparent movement with subjective contours," *Vision Res.*, **13**, 1399-1401.

van Santen, J.P.H. & G. Sperling (1984) "A Temporal Covariance Model of Motion Perception," *J. Opt Soc. Am A*, **1**, 451-473.

van Santen, J.P.H. & G. Sperling (1985) "Elaborated Reichardt Detectors," *J. Opt. Soc. Am. A*, **2**, 2, 300-321.

Shapley, R. & Enroth-Cugell, C., (1984) "Visual adaptation and retinal gain controls", *Prog. Retinal Res.*, **3**, 263-346, 1984.

Sperling, G. (1976) "Movement perception in computer-driven visual displays," *Behavior Research*

Methods and Instrumentation, 8, 144-151.

Sutter, A., Beck, J., & Graham, N. (1989) "Contrast and spatial variables in texture segregation: Testing a simple spatial-frequency channels model," *Perception & Psychophysics*, 46, *4*, 312-332.

Turano, K. & A. Pantle (1989) "On the mechanism that encodes the movement of contrast variations: velocity discrimination," *Vision Res.*, 29, *2*, 207-221.

Ullman, S. (1979) *The Interpretation of Visual Motion.*

Victor, J.D. & M.M. Conte (1990) "Motion mechanisms have only limited access to form information," *Vision Res.*, 30, 2, 289-301.

Watson, A.B. & A.J. Ahumada Jr. (1983a) "A linear motion sensor," *Perception*, 12, A17.

Watson, A.B. & A.J. Ahumada Jr. (1983b) "A Look at Motion in the Frequency Domain," NASA Technical Memorandum 84352.

Watson, A.B. & A.J. Ahumada Jr. (1985) "A model of human visual-motion sensing," *J. Opt. Soc. Am. A*, 2, 2, 322-342.

Watson, A.B., A.J. Ahumada Jr. & J.E. Farrell (1986) "The window of visibility: A psychophysical theory of fidelity in time-sampled motion displays," *J. Opt. Soc. Am A*, 3, *3*, 300-307.

Wilson, J. & S. Sherman, (1976) "Receptive field characteristics of neurones in cat striate cortex: changes with visual field eccentricity," *J. Neurophysiol.*, 39, 512-533.

Figure legends

Fig. 1. The Reichardt detector. Let I be a random stimulus. Then, in response to I, for $i = 1, 2$, the box containing the spatial function $f_i : \mathbf{Z}^2 \rightarrow \mathbf{R}$, outputs the temporal function, $\sum_{(x,y) \in \mathbf{Z}^2} f_i[x, y] I[x, y, t]$; each of the boxes marked $g_i *$ outputs the convolution of its input with the temporal function $g_i : \mathbf{Z} \rightarrow \mathbf{R}$; each of the boxes marked with a multiplication sign outputs the product of its inputs; the box marked with a minus sign outputs its left input minus its right; and the box containing $h*$ outputs the convolution of its input with the temporal function $h : \mathbf{Z} \rightarrow \mathbf{R}$. To see how the Reichardt detector senses motion, suppose f_2 is identical to f_1, but shifted in space by some offset, and suppose the filters $g_1 *$ do not alter their input, while the filters $g_2 *$ simply delay their input by some amount δ_t of time. Then a rigidly translating pattern moving in the direction of box f_2's offset from box f_1 will elicit some time-varying response from box f_1, and the same response a short time later from box f_2. If that "short time later" is precisely δ_t, the output of the righthand multiplier will be positive as long as the pattern keeps drifting. This will result in a net negative Reichardt detector output. If the pattern drift is in the opposite direction, the detector response will be positive.

Fig. 2. Exposing the motion of the traveling contrast-reversal of the random black-or-white vertical bar pattern J to standard motion-analysis. (a) An xt cross-section of J. (b) An xt cross-section of the partial derivative of J with respect to time. (c) An xt cross-section of $|\partial J / \partial t|$. Each of J and $\partial J / \partial t$ is microbalanced However, $|\partial J / \partial t|$ is not. In particular, $|\partial J / \partial t|$ has most of its energy at those frequencies whose velocity is equal to the velocity of the traveling contrast-reversal.

Fig. 3. Fourier and nonFourier motion mechanisms. (a) Fourier motion mechanisms apply standard motion-analysis directly to the luminance signal L. (b, c, d) NonFourier mechanisms apply standard motion analysis to a nonlinear transformation of luminance. (b) A simple nonFourier mechanism applies a signal transformation comprised of a spatiotemporal linear filter, followed by a pointwise nonlinearity. The * 's indicate spatial and temporal convolution, respectively, and • indicates multiplication. The filtering

performed in (b) is roughly pointwise in time (the temporal impulse response b2 approximates an impulse), and the nonlinearity applied is a full-wave rectifier. This system (with appropriately chosen spatial filter, b1) will extract the motion of the texture quilts shown in Figs. 4b, 5d, 6c, and 6d. It will not extract the motion of stimulus J, the traveling contrast-reversal of the random vertical bar pattern shown in Fig. 2a. (c) A spatially pointwise (the spatial impulse response c1 approximates an impulse), system with a flicker-sensitive temporal filter and a full-wave rectifier. Because of the flicker sensitivity, this mechanism will extract the motion of the traveling contrast-reversal of the random vertical bar pattern shown in Fig. 2a but not the motion of the texture quilts shown in Figs. 4b, 5d, 6c, and 6d. (d) The temporal filter d2 averages the temporal filters b2 and c2, and the pointwise nonlinearity is a full-wave rectifier. With an appropriate spatial filter d1, this nonFourier system extracts the motion of any corresponding texture quilt as well as the motion of the traveling contrast-reversal of the random vertical bar pattern shown in Fig. 2a. However, it would be less well-suited to these tasks than the detectors shown in (b) and (c) whose temporal filters it averages.

Fig. 4. Edge-driven motion from an ordinary edge and from a binary texture quilt. (a) A rightward moving light-dark edge visible to Fourier and nonFourier motion systems. Nine entire frames are shown, each frame consists of an area of contrast +1 and area of contrast -1. (b) A realization of the *sidestepping, randomly contrast-reversing vertical edge*. This random stimulus is a texture quilt and hence microbalanced under all purely temporal transformations. that is, its rightward motion would be inaccessible to standard motion analysis even if this analysis were preceded by an arbitrary, purely temporal transformation. Each frame of (b) was derived from the corresponding frame of (a) by multiplying the entire frame by a random variable that takes the value 1 or −1 with equal probability. The frame random variables are jointly independent. A straightforward way to extract the motion of this texture quilt is to (i) apply a linear filter sensitive to vertical edges, (ii) rectify the filtered output, and (iii) submit the result to standard motion analysis.

Fig. 5. Orientation-driven nonFourier motion from a binary texture quilt. (a) A probabilistically defined sinewave grating that steps rightward 90 degrees between frames. The rightward motion in (a) is accessible to all motion detectors. (b1) Four frames of a static, vertical squarewave grating; (b2) Four frames of a static horizontal squarewave grating. (c) A rightward translating texture pattern. For every white point in (a), the corresponding value in (c) is chosen from the vertical square-wave grating in (b1); for every black point in (a), the corresponding value in (c) is chosen from the horizontal square-wave grating in (b2). (c) is not microbalanced; standard motion-analyzers can be designed to detect its motion. (d) A texture quilt. The frames of (d) are derived by multiplying the corresponding frames of (c) by jointly independent random variables, each of which takes the value 1 or −1 with equal probability. The texture quilt (d) is microbalanced under all purely temporal transformations, and therefore its rightward motion is unavailable to any mechanism that applies standard motion analysis to a purely temporal transformation of the visual signal.

Fig. 6. Sinusoidal texture quilts. Motion driven by differences in *orientation* and in *spatial frequency* (b) and (c) show realizations of random stimuli, each of which is microbalanced under all purely temporal transformations. Their rightward motion cannot be detected by any mechanism that applies standard motion analysis to a purely temporal transformation of the signal. In each case, the 4 frames in (a) select between two sinusoidal patterns. The phases of sinusoids are jointly independent across frames and across different-frequency sinusoidal components patched together in the same frame. The sinusoids mixed in (b) differ in orientation, whereas the sinusoids mixed in (c) have the same orientation, but differ in spatial frequency.

Fig. 7. Three quilts used to study motion carried by modulation of texture spatial frequency, by texture orientation, and by higher order textural characteristics. (a) Eight frames that comprise one cycle of the F-quilt. Motion is generated by by a squarewave modulation of textural spatial frequency. The squarewave grating selects between vertical sinusoidal gratings of spatial frequency 1.2 c/deg and 2.4 c/deg The texture-modulating squarewave is 0.3 c/deg, and steps 1/4 cycle rightward on every odd frame. Every even

frame is independent of and distributed identically to the preceding frame. Presentation proceeds at the rate of 30 frames/sec. This gives the texture-modulating squarewave a temporal frequency of 3.75 Hz and a a mean velocity of 25 deg/sec.

(b) Eight frames that comprise once cycle of the O-quilt. In the O-quilt, textural orientation is modulated by the same squarewave used to modulated spatial frequency in the F-quilt. The O-quilt squarewave selects between oppositely oriented sinusoidal gratings that have a spatial frequency of 2.8 c/deg.

(c) Eight frames that comprise once cycle of the E-quilt. In the E-quilt, the texture-modulating squarewave selects between jointly independent binary noise and an "even" texture (Julesz, Gilbert & Victor, 1978) Despite the evident difference between these two textures, every time-independent linear filter has the same expected power for both textures. Thus, if motion-from-texture resulted from applying a simple squaring transformation to the output of a spatial linear filter and submitting the result to standard motion analysis, the motion of the E-quilt would be invisible.

Fig. 8. The percent of correct direction-of-motion judgments to the F-quilt, the O-quilt, and the E-quilt as a function of stimulus duration. The panels show data for subjects CC and GA, respectively. Each data point is the mean of 100 judgments. (Squares) F-quilt; (triangles) O-quilt; (circles) E-quilt. The stimulus durations of 133, 266, 400, and 533 ms, correspond to stimulus presentations of 0.5, 1, 1.5 and 2 quilt cycles.

Anne Sutter, George Sperling and Charles Chubb. Furthur Measurements of the Spatial Frequency Selectivity of Second-Order Texture Mechanisms. Investigative Opthalmology and Visual Science, 1991, 32, No. 4, *ARVO Supplement*, 1039

FURTHER MEASUREMENTS OF THE SPATIAL FREQUENCY SELECTIVITY OF SECOND-ORDER TEXTURE MECHANISMS

Anne Sutter, George Sperling, & Charles Chubb

Human Information Processing Laboratory, New York University, NY, NY 10003

A number of investigations of texture and motion perception suggest a two-stage processing system consisting of an initial stage of selective linear filtering, followed by a rectification and a second stage of selective linear filtering. Here we present new data measuring two properties of the second-stage filters: their contrast modulation sensitivity as a function of spatial frequency (MTF), and the relation of initial spatial filtering to second-stage selectivity. To determine the MTF, we used a staircase procedure to obtain amplitude modulation thresholds for the detection of the orientation of Gabor modulations of a bandlimited noise carrier. We used improved noise carriers with a narrower bandwidth than the stimuli reported last year. Four carrier bands were created with center frequencies of 2, 4, 8, and 16 c/deg. The spatial frequency of the test signals (Gabor amplitude modulations) ranged from 0.5 to 8 c/deg.

The improvements in our stimuli produced a different pattern of results. (1) The threshold amplitude of signal modulation was lowest for 0.5 and 1.0 c/deg. Above 1.0 c/deg. threshold increased with frequency[1]. (2) There was a significant interaction of carrier frequency band with the modulating frequency, with the lowest thresholds occuring for carrier frequency/modulation frequency ratios of about three to four octaves. These results indicate that the second-stage selective filters and detectors are most sensitive to frequencies lower than or equal to 1 c/deg, and that they are selective with regard to the spatial frequency content of the carrier noise on which the signals are impressed.

[1] Jamar, J.H.T. & Koenderink, J.J., (1985). *Vis Res* 25 (4) pp 511-521.

Supported by AFOSR Life Sciences Directorate Grant 88-0140 and NIMH Grant 5T32MH14267

Atch-3

Peter Werkhoven, Charles Chubb and George Sperling. Texture-Defined Motion is Ruled by an Activity Metric--Not by Similarity. Investigative Opthalmology and Visual Science, 1991, 32, No. 4, *ARVO Supplement*, 829

TEXTURE-DEFINED MOTION IS RULED BY AN ACTIVITY METRIC – NOT BY SIMILARITY

Peter Werkhoven, Charles Chubb and George Sperling.

Human Information Processing Laboratory, New York University

We examined motion carried by textural properties. The stimuli we used consisted of patches of sinusoidal grating of various spatial frequencies and contrasts. Phases were randomized to insure that motion mechanisms sensitive to correspondences in stimulus luminance were not systematically engaged.

We used an ambiguous apparent motion paradigm in which a "heterogeneous" motion path (defined by alternating patches of a type A and a type B texture) competes with a "homogeneous" motion path defined by patches of type A. We found that the strength of these (2nd order) motion stimuli is determined by the covariance of the *activity* of the textures that define the motion paths. The activity of a texture is an hypothesized property that is proportional to the texture's contrast and is found to be inversely proportional to its spatial frequency (within the range of spatial frequencies examined). Indeed, heterogeneous motion between equal contrast patches of a high spatial-frequency texture A and a low-spatial frequency texture B can easily dominate homogeneous motion between two patches of A because the activity of texture B is higher than that of texture A.

At temporal frequencies higher than 4 Hz, we find that activity covariance almost exclusively determines motion strength. At lower temporal frequencies, similarity between textures becomes a significant factor as well.

Supported by AFOSR Life Sciences Visual Information Processing Program, Grant 88-0140

Atch-4

Joshua A. Solomon and George Sperling. Can We See 2nd-Order Motion and Texture in the Periphery? Investigative Opthalmology and Visual Science, 1991, 32, No. 4, *ARVO Supplement*, 714

CAN WE SEE 2nd-ORDER MOTION AND TEXTURE IN THE PERIPHERY?

Joshua A Solomon and *George Sperling*.

Human Information Processing Laboratory, New York University

Stimuli. Our 1st-order stimuli are moving sine gratings. Our 2nd-order stimuli are patches of static visual noise, whose contrasts are modulated by moving sine gratings. Neither the spatial orientation nor the direction of motion of these 2nd-order (drift-balanced) stimuli can be detected by analysis of their Fourier domain power spectra. They are invisible to Reichardt and motion-energy detectors.

Method. For these dynamic stimuli, in the fovea, and at 12 deg eccentricity, we measured contrast modulation thresholds as a function of spatial frequency for discrimination of ±45 deg texture slant and for discrimination of direction of motion. Spatial frequency was varied by changing viewing distance.

Results. For sufficiently low spatial frequencies and sufficiently large contrast modulations, all stimuli are visible both foveally and peripherally. For peripherally viewed 1st-order gratings, the highest spatial frequency at which motion or texture discrimination is possible is about 1/4 that at which the corresponding discrimination is possible for foveally viewed gratings. For peripherally viewed 2nd-order gratings, the highest spatial frequencies at which motion or texture discrimination are possible are somewhat less than 1/4 the frequencies of the corresponding foveal discriminations. Thus, as the stimulus moves peripherally, the visual mechanisms that detect 2nd order motion and texture lose sensitivity somewhat faster than the 1st-order mechanisms.

Conclusions. Under certain specific assumptions, our results suggest the following about the neural detectors involved in these discriminations: (1) For both motion and texture, there are more foveal than peripheral detectors at all spatial frequencies. (2) There are more 1st-order than 2nd-order detectors. (3) On the average, foveal detectors respond to higher spatial frequencies than peripheral detectors. (4) The 2nd-order foveal–peripheral spatial frequency difference is somewhat larger than the 1st-order difference.

Supported by AFOSR Life Sciences, Visual Information Processing Program, Grant 88-0140.

Atch-5

Vision Res. Vol. 31, No 7/8, pp. 1399–1415, 1991
Printed in Great Britain.

0042-6989/91 $3.00 + 0.00

OBJECT SPATIAL FREQUENCIES, RETINAL SPATIAL FREQUENCIES, NOISE, AND THE EFFICIENCY OF LETTER DISCRIMINATION

DAVID H. PARISH and GEORGE SPERLING*

Human Information Processing Laboratory, Department of Psychology and Center for Neural Sciences, New York University, NY 10003, U.S.A.

(Received 7 July 1988, in revised form 2 June 1990)

Abstract—To determine which spatial frequencies are most effective for letter identification, and whether this is because letters are objectively more discriminable in these frequency bands or because can utilize the information more efficiently, we studied the 26 upper-case letters of English. Six two-octave wide filters were used to produce spatially filtered letters with 2D-mean frequencies ranging from 0.4 to 20 cycles per letter height. Subjects attempted to identify filtered letters in the presence of identically filtered, added Gaussian noise. The percent of correct letter identifications vs s/n (the root-mean-square ratio of signal to noise power) was determined for each band at four viewing distances ranging over 32:1. Object spatial frequency band and s/n determine *presence of information* in the stimulus; viewing distance determines retinal spatial frequency, and affects only *ability to utilize*. Viewing distance had no effect upon letter discriminability; object spatial frequency, not retinal spatial frequency, determined discriminability. To determine discrimination efficiency, we compared human discrimination to an ideal discriminator. For our two-octave wide bands, s/n performance of humans and of the ideal detector improved with frequency mainly because linear bandwidth increased as a function of frequency. Relative to the ideal detector, human efficiency was 0 in the lowest frequency bands, reached a maximum of 0.42 at 1.5 cycles per object and dropped to about 0.104 in the highest band. Thus, our subjects best extract upper-case letter information from spatial frequencies of 1.5 cycles per object height and they can extract it with equal efficiency over a 32:1 range of retinal frequencies, from 0.074 to more than 2.3 cycles per degree of visual angle.

Spatial filtering Scale invariance Psychophysics Contrast sensitivity Acuity

INTRODUCTION

Characterizing objects

When we view objects, what range of spatial frequencies is critical for recognition, and how is our visual system adapted to perceive these frequencies? Ginsburg (1978, 1980) was among the first to investigate this problem by means of spatial bandpass filtered images of faces and lowpass filtered images of letters. He noted the lowest frequency band for faces and the cutoff frequency for letters at which the images seemed to him to be clearly recognizable. The cutoff frequency for letters was 1–2 cycles per letter width, faces were best recognized in a band centered at 4 cycles per face width. He also proposed that the perception of geometric visual illusions, such as the Mueller–Lyer and Poggendorf, was mediated by low spatial frequencies (Ginsberg, 1971, 1978; Ginsberg & Evans, 1979).

An issue that is related to the lowest frequency band that suffices for recognition is the encoding economy of a band. For a filter with a bandwidth that is proportional to frequency (e.g. a two-octave-wide filter), the lower the frequency, the smaller the number of frequency components needed to encode the filtered image of a constant object. Combining these two notions, Ginsburg concluded that objects were best, or most efficiently, characterized by the lowest band of spatial frequencies that sufficed to discriminate them. Ginsburg (1980) went on to suggest that higher spatial frequencies were redundant for certain tasks, such as face or letter recognition.

Several investigators were quick to point out that objects can be well discriminated in various spatial frequency bands. Fiorentini, Maffei and Sandini (1983) observed that faces were well recognized in either high or in lowpass filtered bands. Norman and Erlich (1987) observed that high spatial frequencies were essential for discrimination between toy tanks in photographs.

*To whom reprint requests should be addressed.

Atch-6

With respect to geometric illusions, both Janez (1984) and Carlson, Moeller and Anderson (1984) observed that the geometric illusions could be perceived for images that had been highpass filtered so that they contained no low spatial frequencies. This suggests that low and high spatial frequency bands may carry equivalently useful information for higher visual processes.

Characterizing the visual system

In the studies cited above, the discussion of spatial filtering focuses on *object* spatial frequencies, that is, frequencies that are defined in terms of some dimension of the object they describe (cycles per object). Most psychophysical research with spatial frequency bands has focused on *retinal* spatial frequencies, that is, frequencies defined in terms of retinal coordinates. For example, the spatial contrast sensitivity function (Davidson, 1968, Campbell & Robson, 1968) describes the threshold sensitivity of the visual system to sine wave gratings as a function of their *retinal* spatial frequency. Visual system sensitivity is greatest at 3–10 cycles per degree of visual angle (c/deg). How does visual system sensitivity relate to object spatial frequencies?

Unconfounding retinal and object spatial frequencies

Retinal spatial frequency and object spatial frequency can be varied independently to determine whether certain object frequencies are best perceived at particular retinal frequencies. Object frequency is manipulated by varying the frequency band of bandpass filtered images, retinal frequency is manipulated by varying the viewing distance.

The cutoff *object* spatial frequency of lowpass filters and the observer's viewing distance were varied independently by Legge, Pelli, Rubin and Schleske (1985) who studied reading rate of filtered text at viewing distances over a 133 1 range. Over about a 6 1 middle range of distances, reading rate was perfectly constant, and it was approximately constant over a 30 1 range. At the longest viewing distances, there was a sharp performance decrease (as the letters became indiscriminably small). At the shortest viewing distance, performance decreased slightly, perhaps due to large eye movements that the subjects would have to execute to bring relevant material towards their lines of sight, and to the impossibility of peripherally previewing new text.

While viewing distance changed the overall level of performance in Legge et al., the cutoff *object* frequency of their low-pass filters at which performance asymptoted did not change. From this study, we learn that reading rate can be quite independent of retinal frequency over a fairly wide range, and that dependence on critical object frequency does not depend on viewing distance. Because the authors measured reading rate only in lowpass filtered images, we cannot infer reading performance in higher spatial frequency bands from their data.

Unconfounding object statistics and visual system properties

Human visual performance is the result of the combined effects of the objectively available information in the stimulus, and the ability of humans to utilize the information. In studying visual performance with differently filtered images, it is critical to separate availability from ability to utilize. For example, narrow-band images can be completely described in terms of a small number of parameters—Fourier coefficients or any other independent descriptors—than wide-band images. Poor human performance with narrow-band images may reflect the impoverished image rather than an intrinsically human characteristic—an ideal observer would exhibit a similar loss.

The problem of assessing the utility of stimulus information becomes acute in comparing human performance in high and in low frequency bandpass filtered images. Typically, filters are constructed to have a bandwidth proportional to frequency (constant bandwidth in terms of octaves). For example, Ginsburg (1980) used faces filtered into 2-octave-wide bands, while Norman and Ehrlich (1987) also used 2-octave bands for their filtered tank pictures. With such filters, high spatial frequency images contain more independent frequencies than low frequency images.

Although linear bandwidth represents perhaps the important difference between images filtered in octave bands at different frequencies, the informational content of the various bands also depends critically on the nature of the specific class of objects, such as faces or letters. Obviously, determining the information content of images is a difficult problem. When it is not solved the amount of stimulus information available within a frequency band is confounded

with the ability of human observers to use the information. Direct comparisons of performance between differently filtered objects are inappropriate. This distinction between objectively available stimulus information and the human ability to use it has not been adequately posed in the context of spatial bandpass filtering.

Efficiency

In the present context, physically available information is best characterized by the performance of an ideal observer. If there were no noise in the stimulus, the ideal observer would invariably respond perfectly. To compare the performance of an observer, human or ideal, noise of root-mean-square (r.m.s.) amplitude n is progressively added to the signal of r.m.s. amplitude s until the performance is reduced to some criterion, such as 50% correct in a letter identification task. This defines the signal to noise ratio, $(s/n)_c$, for a criterion c. Efficiency *eff* of human performance is defined by

$$eff = \left(\frac{s_i}{n_i}\right)_c^2 \Big/ \left(\frac{s_h}{n_h}\right)_c^2$$

where h and i indicate *human* and *ideal* observers, and s and n are r.m.s. signal and noise amplitudes (Tanner & Birdsall, 1958). In a pure, quantally limited system, efficiency actually represents the fraction of quanta absorbed (utilization efficiency). In the context of signal detection theory, efficiency is given by a d' ratio

$$eff = (d_h'/d_i')^2$$

Overview

For an object that contains a broad spectrum of spatial frequencies, object spatial frequency is determined by the center frequency of a spatial bandpass filtered image. Retinal spatial frequency is determined by the viewing distance at which the stimulus is viewed. Stimulus information is determined jointly by the signal-to-noise ratio, by the spatial filtering, and by the characteristics of the set of signals. these three informational components are combined in the efficiency computation. Letters are a convenient stimulus to study because they are highly overlearned so that human performance can be expected to be reasonably efficient, and because much is already known about the visibility of letters in the presence of internal noise (letter acuity) and about the visual processing of letters.

Specifically, to [illegible]e the roles of object and retinal spa[illegible]quencies, letters are filtered into vario[illegible]quency bands. Noise is added, and the p[illegible]ometric function for correct identificatic[illegible] determined as a function of s/n. Accuracy [illegible]ends only on s/n and not on overall contrast, for a wide range of contrasts (Pavel, Sperling, Riedl & Vanderbeck, 1987). This determination is repeated for every combination of object frequency band and viewing distance. Thereby, retinal spatial frequency and object spatial frequency are unconfounded, enabling us to determine whether a particular object frequency band is better discriminated in one visual channel (retinal frequency) than any other (Parish & Sperling, 1987a, b). Moreover, by computing an ideal observer for the identification task, we obtain an objective measure of the information that is present in each of the frequency bands. Finally, the comparison of human performance with the performance of the ideal observer gives us a precise measure of the ability of our subjects to utilize the information in the stimulus. Having untangled these factors, we can determine which spatial frequencies most efficiently characterize letters for identification.

METHOD

Two experiments were conducted using similar stimuli and procedures.

Stimuli

Letters (signals) and noise. The original, unfiltered letters were selected from a simple 5 × 7 upper-case font commonly used on CRT terminals. Since this is an experiment in pattern recognition, we felt that the simplest letter pattern might be the most general, indeed, this font has been widely used in letter discrimination studies. For the purpose of subsequent spatial filtering, the letters were redefined on a pixel grid that measured 45 (vertical height) × 35 (maximum horizontal extent of letters M and W). The letters had value 1 (white), the background had value 0 (black). To avoid edge effects in filtering, the background was extended to 128 × 128 pixels for all computations. However, only the center 90 × 90 pixels of the stimulus were displayed, as these contained effectively all the usable stimulus information, even for low spatial-frequency stimuli. Letters f[illegible] presentation were chosen pseudo-randoml[illegible]m the set of 26 upper-case English letters.

Table 1. Parameters of the bandpass filters lower and upper half-amplitude frequencies, peak, and 2D mean frequencies in cycles/letter height

Band	Lower	Peak	Upper	Mean[a]
0	0	Lowpass	0.53	0 39
1	0 26	0 53	1 05	0 74
2	0 53	1 05	2.11	1.49
3	1 05	2.11	4 22	2 92
4	2 11	4.22	8 44	5 77
5	6.33	Highpass	22.5	20 25

[a]Frequencies are weighted according to their squared amplitude (power) in computing the mean

fields were defined on a 128 × 128 array by choosing independent Gaussian noise samples for each pixel, with the mean equal to zero and a variance σ^2 as required by the condition. (As with the letters, only the central 90 × 90 pixels were displayed.) Forty different noise fields were created.

Filters. Each stimulus consisted of a filtered letter added to an identically filtered noise field. Six spatial filters were available, corresponding to six successive levels of a Laplacian pyramid (Burt & Adelson, 1983). The zero-frequency component was added to the images so that they could be viewed. The object-relative filter characteristics, upper and lower half-amplitude cutoff and 2D mean frequency (cycles per letter height), appear in Table 1. The 2D mean frequency $\bar{f}$ for a given band is

$$\bar{f} = \sum_{x=0}^{127}\sum_{y=0}^{127} f_{x,y}a_{x,y}^2 \Big/ \sum_{x=0}^{127}\sum_{y=0}^{127} a_{x,y}^2,$$

where $f_{x,y}$ is the 2D frequency and $a_{x,y}$ is its amplitude. Cycles per object height is used rather than the more usual cycles per object width because the height of our upper-case letters remained constant across the entire set, whereas the width varied between letters.

The transfer functions (spectra) of the filters are displayed in Fig. 1. Approximately, filters are separated in spatial frequency by an octave (factor of 2) and have a bandwidth at half-amplitude of two octaves. The small mound in the lower right corner of Fig. 1 is a negligible imperfection in filter 4. For convenience, the limited range of spatial frequencies passed by each of the filters will be referred to as the *band* of that filter, a specific band is b_i ($i = 0, 1, 2, 3, 4, 5$), where b_0 is the lowest set of frequencies and b_5 is the highest.

The filter spectra (shown in Fig. 1) are approximately symmetrical in log frequency coordinates; a symmetrical spectrum in log coordinates is highly skewed to the right in linear frequency coordinates, resulting in a mean that

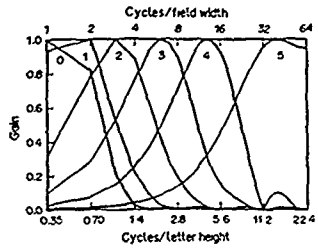


Fig. 1. Filter characteristics for the filters used in the experiments. There are two abscissas, both on a log scale. The top abscissa is the frequency in cycles per unwindowed field width (128 pixels), the bottom abscissa is in cycles per letter height (45 pixels). The ordinate is the normalized gain. The parameter i indicates the filter designation b_i in the text.

is much greater than the mode. In a 2D (vs. 1D) filter, the rightward shift is accentuated. For example, band 2 has a peak frequency of 1.05 c/object but a 2D mean frequency of 1.49 c/object. The single most informative characterization of such a skewed bandpass spectrum depends somewhat on the context, usually use the mean rather than the peak.

Figure 2 (top) shows the letter G, filtered in bands 1–5 without noise; the bottom shows the same signals plus noise. $s/n = 0.5$. The full 128 × 128 array (extended by reflection beyond its edges) was passed through the filter so that the effect of the picture boundary did not intrude into the critical part of the display.

Signal to noise ratio, s/n. A filtered letter is a *signal*. Let i, j index a particular pixel in the x, y coordinate space of the stimulus. The signal contrast $c_s(i, j)$ of pixel i j is

$$c_s(i, j) = \frac{(l(i, j) - l_0)}{l_0} \qquad (1)$$

where $l_{i,j}$ is the luminance of pixel i j and l_0 is the mean signal luminance over the 90 × 90 array. Signal power per pixel, s, is defined as mean contrast power averaged over the 90 × 90 pixel array:

$$s = (IJ)^{-1}\sum_{i}^{I}\sum_{j}^{J} c_s(i, j)^2 \qquad (2)$$

where $c_{i,j}$ is the contrast of pixel i, j and $I = J = 90$.

Noise contrast $c_n(i, j)$ is the value of the i, jth noise sample divided by the mean luminance. Analogously to signal power (equation 2), noise contrast power per pixel, n, is equal to $(\sigma / l_0)^2$. The signal to noise ratio is simply s/n.

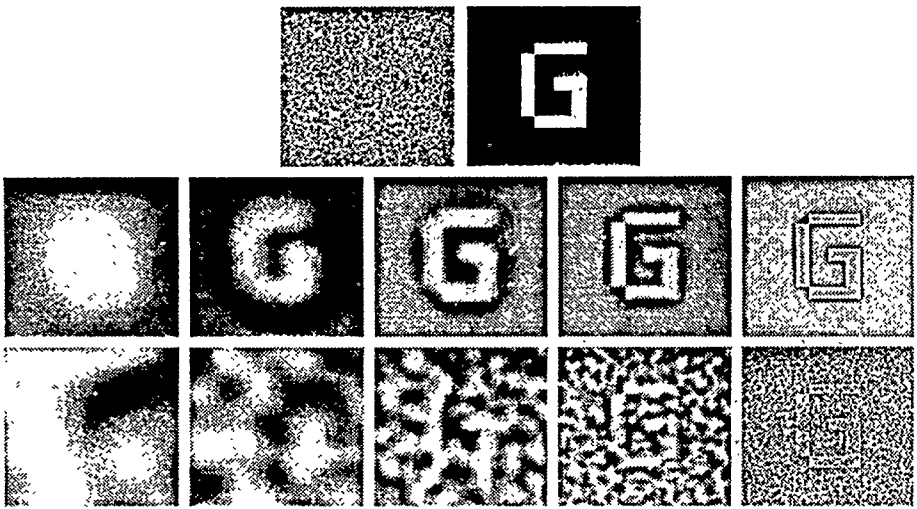


Fig 2 Top unfiltered noise and unfiltered letter G Middle the letter G filtered in spatial frequency bands 1–5 with only quantization noise Bottom filtered letter G plus filter noise in the same bands with a signal-to noise ratio of 0 50 in all panels The effective s/n in the reproduction is somewhat lower (from Parish & Sperling, 1987a) The first row of numerals indicates the number by which the filter band is referred to in the text the bottom row indicates the *mean* frequency of the bands in cycles per letter height

Quantization. Our display system produced 256 discrete luminance levels. Level 128 was used as the mean luminance l_0; l_0 was 47.5 cd/m^2. To produce a visual display of a given letter, band, and s/n, signal power s and noise power n were normalized so that the luminance of every one of the 8100 displayed pixels fell within the range of the display system, there was no truncation of the tails of the Gaussian noise. (Although the relationship between input gray-level and output luminance was not quite linear at the extreme intensity values, it was determined that more than 90% of the pixels fell within the linear intensity range.) Intensity normalization was applied separately to each stimulus (combination of signal plus noise) By normalizing the total stimulus $s+n$, the actual value of s displayed to the subject diminished as n increased, i.e. the actual value of s was not known by the subject. Indeed, even stimuli with precisely the same letter in the same band and with the same s/n might be produced with slightly different s and n depending on the extreme values of the noise fields.

Seven values of s/n were available for each band, chosen in a pilot study to insure that the data yielded the entire psychometric function (chance to best performance). The same pilot study showed that subjects never performed above chance when confronted with noise-free letters from b_0, this band was omitted from the present study

Procedure experiment 1

Four of the experimental variables—letter identity, noise field, frequency band, and s/n—were randomized within each session A fifth variable, viewing distance, was held constant within each session and was varied between sessions Four viewing distances were used. 0 121, 0 38, 1 21 and 3 84 m. A chin rest was used to stabilize the subject's head for viewing at the shortest distance At the four distances, the 90 × 90 pixel stimulus subtended 31.6, 10, 3 16 and 1 0 deg of visual angle respectively. The upper and lower half-amplitude cut-off retinal frequencies for the upper six filters, with respect to the four viewing distances used in this experiment, and for a fifth distance used in the second experiment, appear in Table 2. Subjects participated in four 1-hr sessions at each viewing distance. Each session consisted of 315 trials, nine trials at each of seven s/n's for each of the five frequency bands.

Prior to the first session, subjects were shown noise-free examples of the unfiltered letters They were told that each stimulus presentation consisted of a letter and a certain amount of noise, and that the letter may appear degraded in some way. They were informed that at no time would a letter be shifted in orientation or from its central location in the stimulus field Finally, they were instructed to view each stimulus for as long as they desired before making their best guess as to which letter had been presented. A response (letter identity) was required on every trial. Subjects typed the response on a keyboard connected to the host computer (Vax 11/750); subsequently, typing a carriage return erased the video screen and initiated the next trial in a few seconds The room illumination was very dim, the response keyboard was lighted by stray light from its associated CRT terminal. No feedback was offered to the subjects

Observers

Three subjects, two male and one female, between the ages of 20 and 27 participated in the experiment. All subjects had normal or corrected-to-normal vision One of the subjects was a paid participant in the study

Procedure experiment 2

This experiment was run before expt 1 It is reported here because it offers additional data with two new and one old subject at a fifth viewing distance. Except as noted, the procedures are similar to expt 1 The screen was viewed through a darkened hood at a distance

Table 2 Lower and upper half power frequency and 2D mean frequency (in c/deg of visual angle) for all bands and viewing distances used in both experiments

Band	Viewing distance (m) 0 12	0 38	1 21	3 84	0 45
0 (lowpass)	0 00–0 04 (0 03)	0 00–0 12 (0 09)	0 00–0 37 (0 27)	0 00–1 16 (0 87)	0 00–0 15 (0 11)
1	0 02–0 07 (0 05)	0 06–0 23 (0 16)	0 18–0 74 (0 52)	0 58–2 34 (1 65)	0 07–0 29 (0 21)
2	0 04–0 15 (0 10)	0 12–0 47 (0 33)	0 37–1 48 (1 04)	1 18–4 70 (3 30)	0 15–0 59 (0 41)
3	0 07–0 30 (0 20)	0 23–0 94 (0 64)	0 74–2 97 (2 04)	2 34–9 40 (6 48)	0 29–1 18 (0 81)
4	0 15–0 59 (0 40)	0 47–1 88 (1 27)	1 48–5 94 (4 04)	4 70–18 80 (12 82)	0 59–2 36 (1 60)
5 (highpass)	0 30 2 25 (1 41)	0 94 7 13 (4 45)	2 97–22 53 (14 19)	9 40–71 27 (45 00)	1 77–8 96 (5 63)

of 0.48 m. At this distance, the 90 × 90 stimuli subtended 7.15 deg of visual angle. The half-amplitude cut-off frequencies and the mean frequencies of the six spatial filters are given in the rightmost column of Table 2. Three male subjects between the ages of 20 and 27 participated in the experiment. All subjects had normal or corrected-to-normal vision. Two of the subjects were paid for their participation, and one, DHP, also participated in expt 1. Five sessions of 315 trials were run for each subject.

RESULTS

Psychometric functions: $\hat{p}$ vs $\log_{10} s/n$

The measure of performance is the observed probability $\hat{p}$ of a correct letter identification. The complete psychometric functions are displayed in Figs 3 (expt 1) and 4 (expt 2). A separate psychometric function is shown for each subject, viewing distance and frequency band. In band b_1, for all subjects, performance asymptotes (for noiseless stimuli) at $\hat{p} \approx 0.5$. In all other bands, performance improves from near-chance (1/26) to near perfect as the value of s/n increases.

Noise resistance as a function of frequency band

An obvious aspect of the data of both experiments is that the data move to the left of the figure panels as band spatial frequency increases. This means that high spatial frequency stimuli (bands b_4, b_5) are identifiable at smaller

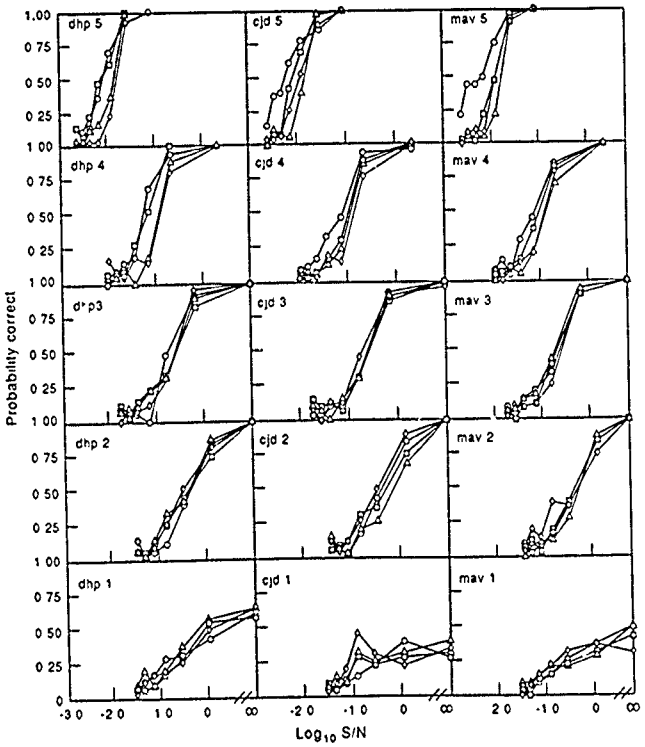


Fig. 3. Psychometric functions from expt 1. Each graph displays performance as a function of $\log_{10} s/n$ within a frequency band. The parameter is viewing distance. Subjects are arranged in columns and frequency band is arranged in rows, progressing from the highest frequency band at the top to the lowest band at the bottom. The four viewing distances are 3.84 (○), 1.21 (△), 0.38 (□), and 0.121 (◇) m.

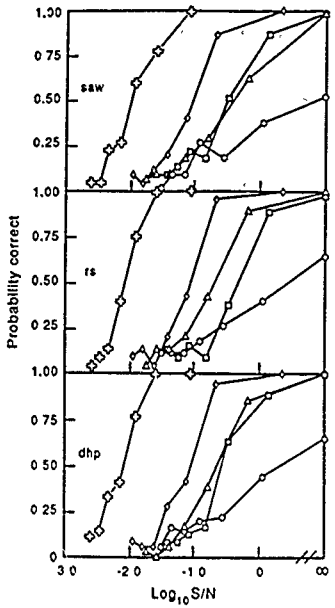


Fig 4 Psychometric functions for each subject and frequency band in expt 2 Viewing distance was 0 48 m The five frequency bands, b_1-b_5, are indicated, respectively, by ○ □, △, ◇ and + The probability of a correct response is plotted as a function of $\log_{10} s/n$

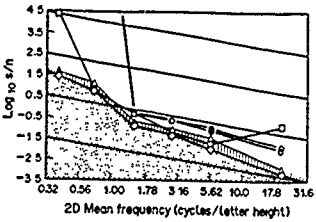


Fig 5 Performance of human subjects and various computational discriminators The abscissa indicates $\log_{10}$ of the mean frequency of each bandpass stimulus The ordinate indicates the (interpolated) s/n ratio at which a probability of a correct response $p = 0.5$ is achieved. Circles indicate each of the three subjects in expt 1 at the intermediate viewing distance of 1.21 m In band b_1, 2 of 3 human subjects fail to achieve 50% correct ($eff = 0$), these points lie outside the graph (△) indicates sub-ideal and (◇) indicates super-ideal performances of discriminators that brackets the ideal discriminator The shaded area below the super-ideal discriminator indicates theoretically unachievable performance Squares indicate performance of a spatial correlator-discriminator The oblique parallel lines have slope −1 that represents the improvement in expected performance (decrease in s/n) as function of the number of frequency components in each band when filter bandwidth is proportional to frequency

s/n than stimuli in bands b_1 and b_2, resistance to noise increases with spatial frequency band. To enable comparisons of noise sensitivity as a function of band, the s/n at which $\hat{p} = 50\%$ was estimated for each subject and frequency band from expt 1 by means of inverse interpolation from the best fitting logistic function As viewing distance had no effect, all estimates were made using the data collected when viewing distance was equal to 0 38 m. A graph of these $(s/n)_{50\%}$, points as a function of the mean object frequency of the band is plotted in Fig 5 (○) For comparison, the expected rate of improvement in $(s/n)_{50\%}$, based on the increasing number of frequency components as one moves from low to high frequency bands, is plotted as a series of parallel lines in Fig 5 Performance improves [$(s/n)_{50\%}$, decreases] somewhat faster than $1/f$ (the slope of the parallel lines) These results, and Fig 5, will be analyzed in detail in the Discussion section

The non-effect of viewing distance

Another property of the data is that, in most conditions, viewing distance has no effect on performance Analysis of variance, carried out individually for each subject, shows that there is no significant effect of distance in any band for subject dhp and a significant effect of distance in bands b_4 and b_5 for the other two subjects Further analysis by a Tukey test (Winer, 1971) in bands b_4 and b_5 for these subjects shows that the only significant effect of distance is that visibility at the longest viewing distance is *better* than at the other three distances For subject CJD, the improvement is equivalent to a gain in s/n of 0 19 and 0 28 $\log_{10}$ (for bands b_4 and b_5, respectively), for MAV, the corresponding gains were 0.21 and 0.40

Improved performance at long viewing distances is almost certainly due to the square configuration of individual pixels, which produces a high frequency spatial pixel noise that is attenuated by viewing from sufficiently far away (Harmon & Julesz, 1973) In low frequency bands, pixel-boundary noise is not a problem because the spatial filtering insures that adjacent pixels vary only slightly in intensity We explored the hypothesis of pixel-boundary noise with subject CJD, who showed a distance effect

in band 5. At an intermediate viewing distance of 1.21 m, CJD squinted her eyes while viewing stimuli from band 5. By blurring the retinal image of the display in this way, performance improved approximately to the level of the furthest viewing distance.

To summarize, the only significant effect of distance that we observed was a lowering of performance at near viewing distances relative to the furthest distance. This impairment occurred primarily in bands 4 and 5. In these bands, the spatial quantization of the display (90 × 90 square-shaped pixels) produces artifactual high spatial frequencies that mask the target. These artifactually produced spatial frequencies can be attenuated by deliberate blurring (squinting), or by producing displays with higher spatial resolution, or by increasing the viewing distance to the point where the pixel boundaries are attenuated by the optics of the eye and neural components of the visual modulation transfer function. In all cases, blurring improves performance and eliminates the slightly deleterious effect of a too small viewing distance. Thus, for correctly constructed stimuli, in the frequency ranges studied, there would be no significant effect of viewing distance on performance. This finding is in agreement with the results of Legge et al. (1985), who examined reading rate rather than letter recognition. It is in stark disagreement with the results of sinewave detection experiments in which retinal frequency is critical—see Sperling (1989) for an explanation.

DISCUSSION

A comparison of performance in different frequency bands shows that subjects perform better the higher the frequency band, and subjects require the smallest signal-to-noise ratio in the highest frequency band. To determine whether performance in high frequency bands is good because humans are more efficient in utilizing high-frequency information, or because there is objectively more information in the high-frequency images, or both, requires an investigation of the performance of an ideal observer. The performance of the ideal observer is the measure of the objective presence of information. Human performance results from the joint effect of the objective presence of information and the ability of humans to utilize that information. Human efficiency is the ratio of human performance to ideal performance.

Ideal discriminator

Definition. An ideal discriminator makes the best possible decision given the available data and the interpretation of "best." The performance of the ideal discriminator defines the objective utility of the information in the stimulus. We prefer the name *ideal discriminator*, rather than *ideal observer*, because it indicates the critical aspect of performance under consideration, but we occasionally use *ideal observer* to emphasize the relations to a large, relevant literature on this subject. Our purposes in this section are first, to derive an ideal discriminator for the letter identification task, second, to develop a practical working approximation to this discriminator, and third, to compare the performance of the human with the ideal discriminator.

Although ideal observers have recently come into greater use in vision research, the applications have focused primarily on determining the limits of performance for relatively low-level visual phenomena. For example, Barlow (1978, 1980), and Barlow and Reeves (1979) investigated the perception of density and of mirror symmetry, Geisler (1984) investigated the limits of acuity and hyperacuity, Legge, Kersten and Burgess (1987) examined the pedestal effect, Kersten (1984) studied the detection of noise patterns; and Pelli (1981) detailed the roles of internal visual noise. Geisler (1989) provides an overview of efficiency computations in early vision. Our application differs from these in that we expand the techniques and apply them to a higher perceptual/cognitive function, letter recognition.

For the letter identification task, the ideal discriminator is conceptually easy to define. A particular observed stimulus, x, representing an unknown letter plus noise, consists of an intensity value (one of 256 possible values) at each of 90 × 90 locations. The discriminator's task is to make the correct choice as frequently as possible from among the 26 alternative letters.

The likelihood of observing stimulus x, given each of the 26 possible signal alternatives, can be computed when the probability density function of the added noise is known exactly. The optimal decision chooses the letter that has the highest likelihood of yielding x. The expected performance of the ideal discriminator is computed by summing its probability of a correct response over the 256^{8100} possible stimuli (256 gray levels, 90 × 90 pixels). Unfortunately,

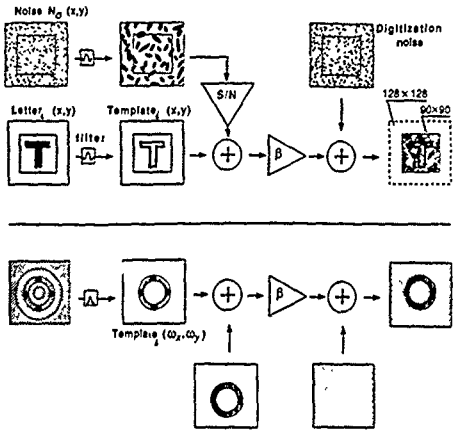


Fig 6 Flow chart of the experimental procedures that are modelled by the ideal discriminator analysis Upper half indicates space domain operations, lower half indicates the corresponding operations in the frequency domain Computations are carried out on 128 × 128 arrays, the subject sees only the center 90 × 90 pixels A random letter and a random noise field are each filtered by the same filter (b), the noise is amplified to provide the desired signal-to noise ratio, the letter and noise are added, the output is scaled and quantized (represented by the addition of digitization noise), and the result is shown to the subject In the frequency domain ω_x, ω_y, the bandpass filter selects an annulus, whereas the quantization noise is uniform over ω_x, ω_y

when there is both bandpass filtered and intensity quantization, the usual simplifications that make this enormous computation tractable are not applicable

As an alternative to computing the expected performance of the ideal discriminator, one can compute its performance with a particular subset of the possible stimuli—the stimuli that the subject actually viewed or, preferably, a larger set of stimuli for more reliable estimation This Monte Carlo simulation of the performance of the ideal discriminator is a tractable computation that yields an estimate of expected performance

Derivation Stimulus construction is diagrammed in Fig 6 which shows the equivalent operations in the space and the frequency domains To derive an ideal discriminator, we need to carefully review the processes of stimulus construction We use uppercase letters to represent quantities in the frequency domain and lowercase letters to represent quantities in the space domain A letter is defined by a 90 × 90 array that takes the value 1 at the letter locations and 0 at the background locations When this array is spatially filtered in band b, it defines the *letter template* $t_{i,b}(x,y)$, where i indicates the particular letter, b the frequency band, and x,y the pixel location We write $T_{i,b}(\omega_x,\omega_y)$ for the Fourier series coefficient of $t_{i,b}$ indexed by frequency

An unknown stimulus $u_{i,b}(x,y)$ to be viewed by a subject is produced by adding filtered $n_b(x,y)$ with post-filtering variance σ_b^2, to the template $t_{i,b}(x,y)$, where letter identity i is unknown to the subject The stimulus is scaled and digitized (quantized) to 256 levels prior to presentation, contributing an additional source of noise $q_{i,b}(x,y)$, called digitization noise Finally, a d.c component (dc) is added to $u_{i,b}$ to bring the mean luminance level to 128 These steps are diagrammed in Fig. 6 which shows both the space-domain and the corresponding frequency-domain operations The space-domain computation is encapsulated in equations (3)

$$u_{i,b}(x,y) = \beta_{i,b}[t_{i,b}(x,y) + n_b(x,y)] \qquad (3a)$$

$$u_{i,b}(x,y) = \beta_{i,b}[t_{i,b}(x,y) + n_b(x,y)] + q_{i,b}(x,y) + dc \qquad (3b)$$

The scaling constant $\beta_{i,b}$ limits the range of real values for each pixel, prior to quantization, to $[-0.5, 255.5]$ The degree of scaling is determined by the maximum and minimum values in

the function $t_{i,k}+n_k$. Note that the extreme values in the image are determined by σ_{λ_2} which is adjusted to yield the appropriate s/n for each condition, the values of $t_{i,k}$ are fixed prior to scaling. Specifically:

$$\beta_{i,k}=\frac{256}{\max(t_{i,k}+n_k)-\min(t_{i,k}+n_k)}. \quad (4)$$

As a result of bandpass filtering, the noise samples in adjacent pixels are strongly dependent on each other. Therefore, the discriminator problem is best approached in the Fourier domain, where the random variables $\{N_k(\omega_x,\omega_y)\}$ are jointly independent because the filtering operations simply scale the different frequency components without introducing any correlations (van Tress, 1968). The task of the ideal discriminator is to pick the template $t_{i,k}$ that maximizes the likelihood of $u_{i,k}$ with *a priori* knowledge of: (i) the fixed functions $t_{i,k}$, and their probabilities, and (ii) the densities of the jointly independent random variables $\{N_k(\omega_x,\omega_y)\}$. As is clear, $\beta_{i,k}$, σ_λ^2, $\{Q_{i,k}(\omega_x,\omega_y)\}$, and $\{N_{i,k}(\omega_x,\omega_y)\}$ are all jointly distributed random variables characterized by some density f. To compute the likelihood of $u_{i,k}$ the ideal discriminator must integrate f over all possible values that may be assumed by the set of jointly distributed random variables, whose values are constrained only in that they result in a possible stimulus $u_{i,k}$. Unfortunately, no closed-form solution to this problem is available, forcing us to look for an alternative approach.

Bracketing. To estimate the performance of the ideal discriminator, we look for a tractable super-ideal discriminator that is better than the ideal but which is solvable. Similarly, we look for a tractable sub-ideal discriminator that is worse than the ideal. The ideal discriminator must lie between these two discriminators, that is, we bracket its performance between that of a "super-ideal" and a "sub-ideal" discriminator. The more similar the performance of the super- and sub-ideal discriminators, the more constrained is the ideal performance which lies between them.

Our super-ideal discriminator is told, *a priori*, the exact values for $\beta_{i,k}$ and σ_λ^2 for each stimulus presentation. Therefore, it is expected to perform slightly better than the ideal discriminator which must estimate these values from the data. The sub-ideal discriminator estimates these same parameters from the presented stimulus in a simple but nonideal way. Therefore, it is expected to perform slightly worse than the ideal discriminator. The computational forms used to compute $\beta_{i,k}$ and σ_λ^2 for the sub-ideal discriminator are presented in the Appendix, along with the derivation of the likelihood estimator used by both discriminators. A complete discussion of these derivations and the problems associated with the formulation of an ideal discriminator for such complex stimuli is presented in Chubb, Sperling and Parish (1987).

Performance of the bracketed discriminator. The super- and sub-ideal discriminators were tested in a Monte Carlo series of trials, in which they each were confronted with 90 stimuli in each of the frequency bands at each of seven s/n values chosen to best estimate their 50% performance point. The s/n necessary for 50% correct discriminations was estimated by an inverse interpolation of the best fitting logistic function. The derived $(s/n)_{50\%}$, is the measure of performance of a discriminator. The mean ratio, across frequency bands, of

$$(s/n)_{50\%}\text{, sub-ideal}/(s/n)_{50\%}\text{, super-ideal}$$

is about 2 (approx. 0.3 $\log_{10}$ units). The ratio does not depend on the criterion of performance.

Efficiency of human discrimination.

In all conditions, human subjects perform worse than the sub-ideal discriminator. Notably, with no added luminance noise, the subideal (and, of course, the ideal) discriminator function perfectly, even in b_0 where subject performance is at chance, and in b_1 where subjects reached asymptote at about 50% correct.

Data from the subjects are plotted with the $(s/n)_{50\%}$, sub-ideal and $(s/n)_{50\%}$, super-ideal in Fig. 5. For comparison, Fig. 5 also shows the performance of a correlator discriminator which chooses the letter template that correlates most highly with the stimulus in the space domain. In the coordinates of Fig. 5 ($\log_{10} s/n$ vs $\log_{10} f$ where f represents the mean 2D spatial frequency of the band), the vertical distance d from the human data $\log(s/n)_{50\%}$, *human* down to the bracketed discriminator $\log(s/n)_{50\%}$, *ideal* represents the $\log_{10}$ of the factor by which the bracketed discriminator outperforms the human observer at that value of f. For the purpose of specifying efficiency, we assume the ideal discriminator lies at the mid-point of the sub and super-ideal discriminators in Fig. 5. The

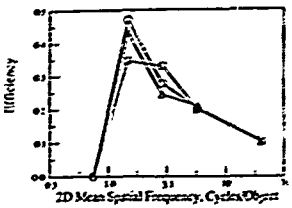


Fig. 7. Discrimination efficiency as a function of the mean frequency of a 2-octave band (in cycles per letter height) indicated on a logarithmic scale. Data are shown for three observers: △ = SAW, □ = RS, ○ = DHP. The viewing distance is 2.21 m, which is representative of all viewing distances tested.

efficiency *eff* of human discrimination relative to the bracketed discriminator is $eff = 10^{-2d}$, where,

$$d = \log(s/n)_{80\%,\ human} - \log(s/n)_{80\%,\ ideal}.$$

The values of *eff* in each object frequency band are shown in Fig. 7. In band 0, *eff* is zero because human performance never reaches 50%, indeed, it never rises significantly above 4% (chance). In band 1, human performance asymptotically climbs close to 50% as s/n approaches infinity, $eff \approx 0$. In band 2, *eff* reaches its maximum of 35–47% (depending on the subject) and it declines rapidly with increasing frequency (b_3–b_5).

The 42% average efficiency in band 2 is similar in magnitude to the highest efficiencies observed in comparable studies. For example, efficiency has been determined for detecting various kinds of patterns in arrays of random dots (Barlow, 1978, 1980, van Meeteren & Barlow, 1981), tasks which, like ours, may require significantly cognitive processing. In a wide range of conditions, the highest efficiencies observed were about 50%, and frequently lower. Van Meeteren and Barlow (1981) also found that efficiency was perfectly correlated with object spatial frequency and was independent of retinal spatial frequency.

Spatial correlator discriminator. A correlator discriminator cross-correlates the presented stimulus with its memory templates and chooses the template with the highest correlation. Correlation can be carried out in the space or in the frequency domain. Correlation is an efficient strategy when noise in adjacent pixels is independent and when members of the set of signals have the same energy, both of these conditions are violated by our stimuli. However, when sufficient prior information is available to subjects, they do appear to employ a cross-correlation strategy (Burgess, 1985).

It is interesting to note that the performance of the spatial correlator discriminator over the middle range of spatial frequencies is quite close to the performance of the sub-ideal discriminator. At high spatial frequencies, correlator performance degenerates, due to its inability to focus spatially on those pixel locations that contain the most information. A spatial correlator that optimally weighted spatial locations, could overcome the spatial focusing problem at high frequencies. (Spatial focusing is treated in the next section.)

At all frequencies, the spatial correlator is nonideal because noise at spatial adjacent pixels is not independent. At low spatial frequencies, the nonindependence of adjacent locations becomes extreme and the correlator fails miserably. This points out that, for our stimuli, correlation detection is better carried out in the frequency domain because there the noise at different frequencies is independent. The qualitative similarity between the correlator discriminator and the subjects' data suggests that the subjects might be employing a spatial correlation strategy, augmented by location weighting at high frequencies.

Lowest spatial frequencies sufficient for letter discrimination. Band 2 corresponds to a 2-octave band with a peak frequency of 1.05 c/object (vertical height of letters) and a 2D mean frequency of 1.49 c/object. At the four viewing distances, 1.05 c/object corresponds to retinal frequencies of 0.074, 0.234, 0.739 and 2.34 c/deg of visual angle. We observe perfect scale invariance: all of these retinal frequencies, and hence the visual channels that process this information, are equally effective in achieving the high efficiency of discrimination.

The finding that b_2 with a center frequency of 1.05 c/object and a $\frac{1}{2}$ amplitude cutoff at 2.1 c/object is critical for letter discrimination is in good agreement with previous findings of both Ginsburg (1978) for letter recognition and Legge et al. (1985) for reading rate. Legge et al. used low-pass filtered stimuli, which included not only spatial frequencies within an octave of 1 c/object (b_2) but also included all lower frequencies. From the present study, we expect human performance with low-pass and with band-pass spatial filtering to be quite similar up to 1 c/object because the lowest frequency

bands, when presented in isolation, are perceptually useless (at least when presented alone).

It is an important fact that our subjects actually performed better, in the sense of achieving criterion performance at a lower s/n ratio, at higher frequency bands than b_2. This is explained by the increase in stimulus information in higher frequency stimuli. Increased information more than compensates for the subjects' loss in efficiency as spatial frequency increases.

Components of discrimination performance

Though the performance of the bracketed ideal discriminator is useful in quantifying the informational utility of the various bands, it is instructive to consider the changing physical structure of the stimuli as well. What components of the stimuli actually lead to a gain in information with increasing frequency? According to Shannon's theorem (Shannon & Weaver, 1949), an absolutely bandlimited 1-D signal can be represented by a number of samples m that is proportional to its bandwidth. When the signal-to-noise ratio in each sample s, n, is the same, the overall signal-to-noise ratio s/n grows as $\sqrt{m}$. In the space domain, our filters were constructed (approximately) to differ only in scale but not in the shape of their impulse responses. Therefore, when the mean frequency of a filter band increased by a factor of 2, the bandwidth also increased by 2. Since the stimuli are 2D, the effective number of samples increases with the square of frequency, and the increase in effective s n ratio is proportional to m. This expected improvement with frequency, based simply on the increase in effective number of samples, is indicated by the oblique parallel lines of Fig. 5 with slope of -1. The expected improvement in threshold s n due simply to the linearly increasing bandwidth of the bands does a reasonable job of accounting for the improvement in performance for both human and bracketed discriminators between b_2 and b_5.

Performance of all discriminators improves faster with frequency between 0.39 and 1.5 c/object and between 5.8 and 22 c/object than is predicted from the bandwidths of the images. A slope steeper than -1 means that there is more information for discriminating letters in higher frequency bands even when the number of independent samples is kept the same in each band. Once sampling density is controlled, just how much information letters happen to contain in each frequency band is an ecological property of upper-case letters.

Increasing spatial localization with increasing frequency band. From the human observer's point of view, the letter information in low-pass filtered images is spread out over a large portion of the total image array. In high spatial-frequency images, the letter information is concentrated in a small proportion of the total number of pixels. In high spatial-frequency images, a human observer who knows which pixels to attend will experience an effective s/n that is higher than an observer who attends equally to all pixels. In this respect, humans differ from an ideal discriminator. The ideal discriminator has unlimited memory and processing resources, does not explicitly incorporate any selective mechanism into its decision, and uses the same algorithm in all frequency bands. Information from irrelevant pixels is enmeshed in the computation but cancels out perfectly in the letter-decision process. To understand human performance, however, it is useful to examine how, with our size-scaled spatial filters, letter information comes to be occupy a smaller and smaller fraction of the image array as spatial frequency increases.

Here we consider three formulations of the change in the internal structure of the images with increasing spatial frequency: (1) spatial localization, (2) correlation between signals, and (3) nearest neighbor analysis. We have already noted that, in our images, the information-rich pixels become a smaller fraction of the total pixels as frequency band increases. Indeed, this reduction can be estimated by computing the information transmitted at any particular pixel location or, more appropriately for estimating noise resistance, by computing the variance of intensity (at that pixel location) over the set of 26 alternative signals.

To demonstrate the degree of increasing localization with increasing frequency, the variance (over the set of 26 letter templates) was computed at each pixel location (x, y). *Total power*, the total variance, is obtained by summing over pixel locations. The number of pixel locations needed to achieve a specific fraction of the total power is given in Fig. 8, with frequency band as a parameter. These curves describe the spatial distribution of information in the letter templates. If all pixels were equally informative, exactly half of the total number of pixels would be needed to account for 50% of the total power. The solid curves in Fig. 8 show that the number of pixels needed to convey any percentage of total signal power, decreases as the

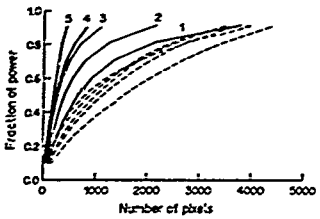


Fig. 8. Fraction of total power contained in the n most extreme-valued pixels as a function of n (out of 8100). Solid lines indicate the power fractions for signals; the curve parameter indicates the filter band. Dashed lines indicate power fractions for filtered noise fields. Although power fractions from successive bands of noise are too close to label, they generally fall in the same left–right 5–0 order as those for signal bands.

frequency band increases. These information distribution curves are an ecological property of our set of letter stimuli; different curves would be needed describe other stimulus sets.

The dashed curves in Fig. 8 were derived from random noise filtered in each of the six frequency bands (b_0–b_5). The distribution of noise power is very similar between the various bands, enormously more so than the distribution of signal power. For our letter stimuli, stimulus information coalesces to a smaller number of spatial locations as spatial frequency increases.

Correlation between signals A more abstract way of describing the change of information with bandwidth is to note that letters become less confusible with each other in the higher frequency bands. A good measure of confusibility is the average pairwise correlation between the 26 letter templates in each frequency band (Table 3). The average correlation between letter templates diminishes from 0.94 in band 0 to 0.31 in band 5. In a band in which templates have a pairwise correlation over 0.9, the overwhelming amount of intensity variation ("information") is useless for discrimination. Small wonder that subjects fail completely in this band. Overall, performance of the ideal discriminator and of observers improves as the correlation decreases, but there is no obvious way to use the pairwise correlation between templates to predict performance.

Nearest neighbors The analysis of nearest neighbors is a useful technique for predicting accuracy by the analysis of the possible causes of errors. We can regard a filtered image t_i of letter i as a vector in a space of dimensionality 8100 (90 × 90 pixels). When noise is added, the possible positions of t_i are described by a cloud whose dimensions are determined by the s/n ratio. A neighboring letter k may be confused with letter i when the cloud around t_i envelopes t_k. The closer the neighbor, the greater the opportunity for error. Table 3 gives the average normalized distance to the nearest neighbor in each of the bands. The increase in distance to the nearest neighbor reflects the improvement in the representation of signals as spatial frequency increases.

Table 3. Average pairwise correlations and nearest neighbors (Euclidean distance × 10^{-3})

Band	Correlations	Nearest neighbor
0	0.94	0.01
1	0.91	0.30
2	0.58	1.2
3	0.38	2.3
4	0.33	3.1
5	0.31	4.1

We consider possible causes of lower efficiency of discrimination in bands below b_2. The letters in these bands have high pair-wise correlations and the mean band frequency is less than the object frequency. This means that letters differ only in subtle differences of shading, a feature that we usually do not think of as shape. Observers would need to be able to utilize small intensity differences to distinguish between letters. To eliminate an alternative explanation (the smaller number of frequency components in the low-frequency bands), we conducted an informal experiment with a lower fundamental frequency. The fundamental frequency, which is outside the band, nevertheless determines the spacing of frequency components within the band. Reducing the fundamental frequency of the letter by one-half increases the number of frequency components in the band by a factor of 4. (A 256 × 256 sampling grid was used rather than 128 × 128.) These 4× more highly sampled stimuli were not more discriminable than the original stimuli. This suggests that the internal letter representation (template) that subjects bring with them to the experiment cannot utilize low-frequency information, even when it is abundantly available. Whether, with sufficient training, subjects could learn to use low spatial frequencies to make letter discriminations is an open question.

SUMMARY AND CONCLUSIONS

1. Visual discrimination of letters in noise, spatially filtered in 2-octave wide bands, is

independent of viewing distance (retinal frequency) but improves as spatial frequency increases.

2. The improvement in performance with increasing spatial frequency results mainly from an increase in the objective amount of information transmitted by the filters with increasing frequency (because filter bandwidth was proportional to center frequency) which is manifested as objectively less confusible stimuli in the higher bands.

3. The comparison of human performance with that of an estimated ideal discriminator demonstrates that humans achieve optimal discrimination (a remarkable 42% efficiency) when letters are defined by a 2-octave band of spatial frequencies centered at 1 cycle per letter height (mean frequency 1.5 c/letter). This high efficiency of discrimination is maintained over a 32:1 range of viewing distances.

4. Detection efficiency was invariant over a range of retinal spatial frequencies in which the contrast threshold for detection of sine gratings (the modulation transfer function, MTF) varies enormously. The independence of detection performance and retinal size held for all frequency bands.

5. A part of the loss of human efficiency in discrimination as spatial frequency exceeded 1 c/object height may have been due to the subjects' inability to identify, to selectively attend, and to utilize the smaller fraction of information-rich pixels in the higher frequency images.

6. Finally, it is important to note that without the comparison to the ideal observer, we would not have been able to understand the components of human performance in the different frequency bands.

Acknowledgements—We acknowledge the large contribution of Charles Chubb to the formulation and solution of the ideal discriminator. We thank Michael S. Landy for helpful comments and Robert Picardi for skillful technical assistance. The project was supported by USAF, Life Sciences Directorate, Visual Information Processing Program, grants 85-0364 and 88-0140.

REFERENCES

Barlow, H. B. (1978). The efficiency of detecting changes of density in random dot patterns. *Vision Research, 18*, 637–650.

Barlow, H. M. (1980). The absolute efficiency of perceptual decisions. *Philosophical Transactions of the Royal Society, London B, 290*, 71–82.

Barlow, H. B. & Reeves, B. C. (1979). The versatility and absolute efficiency of detecting mirror symmetry in random dot displays. *Vision Research, 19*, 783–793.

Burgess, A. (1985). Visual signal detection—III. On Bayesian use of prior knowledge and cross correlation. *Journal of the Optical Society of America A, 2(9)*, 1498–1507.

Burgess, A. (1986). Induced internal noise in visual decision tasks. *Journal of the Optical Society of America A, 3*, 93.

Burt, P. J. & Adelson, E. H. (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications, Com-31(4)*, 532–540.

Campbell, F. W. & Robson, J. G. (1968). Application of Fourier analysis to the visibility of gratings. *Journal of Physiology, London 197*, 551–566.

Carlson, C. R., Moeller, J. R. & Anderson, C. H. (1984). Visual illusions without low spatial frequencies. *Vision Research, 24*, 1407–1413.

Chubb, C., Sperling, G. & Parish, D. H. (1987). Designing psychophysical discrimination tasks for which ideal performance is computationally tractable. Unpublished manuscript, New York University, Human Information Processing Laboratory.

Davidson, M. L. (1968). Perturbation approach to spatial brightness interaction in human vision. *Journal of the Optical Society of America A, 58*, 1300–1309.

Fiorentini, A., Maffei, L. & Sandini, G. (1983). The role of high spatial frequencies in face perception. *Perception, 12*, 195–201.

Geisler, W. S. (1984). Physical limits of acuity and hyperacuity. *Journal of the Optical Society of America A, 1*, 775–782.

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review, 21*, 267–314.

Ginsburg, A. P. (1971). Psychological correlates of a model of the human visual system. In *Proceedings of the National Aerospace Electronics Conference (NAECON)* (pp. 283–290). Ohio: IEEE Trans. Aerospace Electronic Systems.

Ginsburg, A. P. (1978). Visual information processing based on spatial filters constrained by biological data. Aerospace Medical Research Laboratory, 1(2). Dayton, Ohio.

Ginsburg, A. P. (1980). Specifying relevant spatial information for image evaluation and display designs. An explanation of how we see certain objects. *Proceedings of SID, 21*, 219–227.

Ginsberg, A. P. & Evans, P. W. (1979). Predicting visual illusions from filtered imaged based on biological data. *Journal of the Optical Society of America A, 69*, 1443.

Harmon, L. D. & Julesz, B. (1973). Masking in visual recognition: Effects of two-dimensional filtered noise. *Science, 180*, 1194–1197.

Janez, L. (1984). Visual grouping without low spatial frequencies. *Vision Research, 24*, 271–274.

Kersten, D. (1984). Spatial summation in visual noise. *Vision Research, 24*, 1977–1990.

Legge, G. E., Pelli, D. G., Rubin, G. S. & Schleske, M. M. (1985). Psychophysics of reading—I. Normal vision. *Vision Research, 25(2)*, 239–252.

Legge, G. E., Kersten, D. & Burgess, A. E. (1987). Contrast discrimination in noise. *Journal of the Optical Society of America A, 4(2)*, 391–404.

van Meeteren, A. & Barlow, H. B. (1981). The statistical efficiency for detecting sinusoidal modulation of average dot density in random figures. *Vision Research, 21*, 765–777.

van Nes, F. L. & Bouman, M. A. (1967). Spatial modulation transfer in the human eye. *Journal of the Optical Society of America, 57*, 401–406.

Norman, J. & Ehrlich, S. (1987). Spatial frequency filtering and target identification. *Vision Research, 27(1),* 97–96.
Parish, D. H. & Sperling, G (1987a). Object spatial frequencies, retinal spatial frequencies, and the efficiency of letter discrimination. Mathematical Studies in Perception and Cognition, 87–8. New York University, Department of Psychology.
Parish, D. H & Sperling, G (1987b) Object spatial frequency, not retinal spatial frequency, determines identification efficiency. *Investigative Ophthalmology and Visual Science (ARVO Suppl), 28(3),* 359.
Pavel, M , Sperling, G., Riedl, T. & Vanderbeek, A (1987). The limits of visual communication The effect of signal-to-noise ratio on the intelligibility of American sign language *Journal of the Optical Society of America A, 4,* 2355–2365
Pelli, D. G (1981) Effects of visual noise Ph D. dissertation, University of Cambridge, England
Shannon, C. E. & Weaver, W. (1949) *The mathematical theory of communication* Urbana University of Illinois Press
Sperling, G (1989) Three stages and two systems of visual processing *Spatial Vision, 4 (Prazdny Memorial Issue),* 183–207
Sperling G & Parish, D H (1985) Forest-in-the-Trees illusions *Investigative Ophthalmology and Visual Science (ARVO Suppl), 26,* 285
Tanner W P & Birdsall, T G (1958) Definitions of d' and η as psychophysical measures *Journal of the Acoustical Society of America, 30,* 922–928
van Tress, H L (1968) *Detection, estimation and modulation theory* New York Wiley
Winer B J (1971) *Statistical principles in experimental psychology* New York McGraw-Hill

APPENDIX

Both sub-ideal and super-ideal discriminators must compute estimates of the likelihood that the stimulus $u_{i,b}$ was produced with template $t_{k,b}$ and noise n_b where k is the letter used to generate the stimulus, i is an arbitrary letter, and b indexes spatial frequency band Let x be an index on the pixels of the image $1 \leqslant x \leqslant 8100$, for the 90×90 images of the experiments

For the Monte Carlo simulations of the super-ideal discriminator, the unknown stimulus parameters, $\alpha_{i,b}$ and σ_b^2 are computed during stimulus construction and their exact values are supplied to the discriminator *a priori* The sub-ideal discriminator, however must estimate these parameters from the data as follows

Sub-Ideal Parameter Estimation

Recall that stimulus contrast is modulated for any pixel x in the image

$$u_{i,b}[x] = \beta_{i,b}[t_{i,b}(x) + n_b(x)] + q_{i,b}(x) \qquad \text{(A1)}$$

The scaling constant $\beta_{i,b}$ limits range of real values for each pixel, prior to quantization to the open interval $(-0.5, 255.5)$ the addition of $q_{i,b}(x)$ called quantization noise, rounds off pixel values to integers

For each bandpass filtered template $t_{i,b}$ we first compute the correlation $\rho_{k,i}$ of the template to the stimulus $u_{i,b}$

$$\rho_{k,i} = \frac{\sum_x u_{i,b}(x) t_{i,b}(x)}{\left\{\sum_x [u_{i,b}(x)]\right\}^{1/2} \left\{\sum_x [t_{i,b}(x)]^2\right\}^{1/2}} \qquad \text{(A2)}$$

To compute the likelihood estimates for each template $t_{i,b}$, we must be able to reverse the effect of $\beta_{i,b}$. Thus we define $\alpha_{i,b} = 1/\beta_{i,b}$ and choose $\alpha_{i,b}$ so as to minimize the expression

$$\sum_x [\alpha_{i,b} u_{i,b}(x)]^2 = \sum_x [\rho_{k,i} t_{i,b}(x)]^2. \qquad \text{(A3)}$$

Solving for $\alpha_{i,b}$ gives us.

$$\alpha_{i,b} = \rho_{k,i} \left\{ \frac{\sum_x [t_{i,b}(x)]^2}{\sum_x [u_{i,b}(x)]^2} \right\}^{1/2} \qquad \text{(A4)}$$

Finally we set:

$$\sigma_N^2 = \frac{1}{X} \sum_{x=1}^{X} [\alpha_{i,b} u_{i,b}(x) - t_{i,b}(x)]^2 \qquad \text{(A5)}$$

where $X = 8100$, the number of pixels in the image

Likelihood Estimation

With estimates of σ_b^2 and $\alpha_{i,b}$ for the sub-ideal discriminator, and the *a priori* values for the super-ideal discriminator, we can formulate a maximum likelihood estimator By rearranging terms of equation (A1) and dividing both sides by β yields

$$\frac{u_{i,b}(x)}{\beta} - t_{i,b}(x) = n_b(x) + \frac{q_{i,b}(x)}{\beta} \qquad \text{(A6)}$$

Substituting $\alpha_{i,b}$ for $1/\beta$, and by transposing into the frequency domain, denoted by upper-case letters and indexed by ω, we have

$$\alpha_{i,b} U_{i,b}(\omega) - T_{i,b}(\omega) = N_b(\omega) + \alpha_{i,b} Q_{i,b}(\omega) \qquad \text{(A7)}$$

Note that the left side of equation (A7) is simply a difference image between the stimulus $U_{i,b}(\omega)$ and the template $T_{i,b}(\omega)$ This difference is exactly equal to the sum of the luminance and quantization noise only when the correct template is chosen ($i = k$) When the incorrect template is chosen ($i \neq k$) the right hand side of equation (A7) is equal to the sum of the noise sources plus some residue that is equal to $T_{k,b}(\omega) - T_{i,b}(\omega)$ Under the assumption that quantization noise can be modeled as independent additive noise in the frequency domain, the density A of the joint realization of the right-hand side of equation (A7) is given by

$$A = \prod_{\omega} \frac{X}{\pi[\sigma_b^2 \alpha_{i,b}^2 + \sigma_Q^2 |F_b(\omega)|]^2} \times \exp\left[\frac{-X|\alpha_{i,b} U_{k,b}(\omega) - T_{i,b}(\omega)|^2}{\alpha_{i,b}^2 \sigma_b^2 + \sigma_Q^2 |F_b(\omega)|^2}\right] \qquad \text{(A8)}$$

where $F_b(\omega)$ is simply the kernel of filter b, in the frequency domain Dropping the multiplicative term in equation (A8), which does not depend on the template T, and taking logs the ideal discriminator chooses the template that minimizes

$$\sum_{\omega} \frac{X|\alpha_{i,b} U_{k,b}(\omega) - T_{i,b}(\omega)|^2}{\alpha_{i,b}^2 \sigma_b^2 + \sigma_Q^2 |F_b(\omega)|^2} \qquad \text{(A9)}$$

Finally, it is more convenient to compute the power of the quantization noise in the space domain (σ_q^2) than in the frequency domain (σ_Q^2) $\sigma_q^2 = \sigma_Q^2$ Spatial quantization noise, $q_{i,b}(x)$, is uniformly distributed on the interval $[-0.5, 0.5)$ so that σ_q^2 is computed as

$$\int_{-0.5}^{0.5} x^2 \, dx \qquad \text{(A10)}$$

and is equal to 1/12

Vision Res Vol 31, No 5, pp 859–876, 1991
Printed in Great Britain.

0042-6989/91 $3.00 + 0.00

THE KINETIC DEPTH EFFECT AND OPTIC FLOW—II. FIRST- AND SECOND-ORDER MOTION

MICHAEL S. LANDY,[1] BARBARA A. DOSHER,[2] GEORGE SPERLING[1] and MARK E. PERKINS[1]
[1]Psychology Department, New York University, NY 10003 and [2]Psychology Department, Columbia University, NY 10027, U.S.A

(*Received* 24 *August* 1989, *in revised form* 1 *May* 1990)

Abstract—We use a difficult shape identification task to analyze how humans extract 3D surface structure from dynamic 2D stimuli—the kinetic depth effect (KDE). Stimuli composed of luminous tokens moving on a less luminous background yield accurate 3D shape identification regardless of the particular token used (either dots, lines, or disks). These displays stimulate both the 1st-order (Fourier-energy) motion detectors and 2nd-order (nonFourier) motion detectors. To determine which system supports KDE, we employ stimulus manipulations that weaken or distort 1st-order motion energy (e.g. frame-to-frame alternation of the contrast polarity of tokens) and manipulations that create *microbalanced* stimuli which have no useful 1st-order motion energy. All manipulations that impair 1st-order motion energy correspondingly impair 3D shape identification. In certain cases, 2nd-order motion could support limited KDE, but it was not robust and was of low spatial resolution. We conclude that 1st-order motion detectors are the primary input to the kinetic depth system. To determine minimal conditions for KDE we use a two frame display. Under optimal conditions KDE supports shape identification performance at 63–94% of full-rotation displays (where baseline is 5%). Increasing the amount of 3D rotation portrayed or introducing a blank inter-stimulus interval impairs performance. Together, our results confirm that the human KDE computation of surface shape uses a global optic flow computed primarily by 1st-order motion detectors with minor 2nd-order inputs. Accurate 3D shape identification requires only two views and therefore does not require knowledge of acceleration.

KDE Kinetic depth effect Structure from motion Shape Optic flow

INTRODUCTION

When a collection of randomly positioned dots moves on a CRT screen with motion paths that are projections of rigid 3D motion, a human viewer perceives a striking impression of three-dimensionality and depth. This phenomenon of depth computed from relative motion cues is known as the kinetic depth effect (KDE. Wallach & O'Connell, 1953).

What are the important cues that lead to a 3D percept from such a display? Is it motion, or are there other important cues? If it is motion, then what kind of motion detection system(s) are used to support the structure-from-motion computation? Is a computation of velocity sufficient, or are more elaborate measurements necessary, such as of acceleration? These are the questions that we address in this paper.

In a series of recent papers (Dosher, Landy & Sperling, 1989a, b, Sperling, Landy, Dosher & Perkins, 1989, Sperling, Dosher & Landy, 1990), we examined the cues necessary for subjects to perceive an accurate representation of a 3D surface portrayed using random dot displays. In each trial of a new shape identification task we devised, subjects view a random dot representation of one of a set of 52 3D shapes and identify the shape and rotation direction. Shape identity feedback optimizes the subject's ability to compute shape from each type of motion stimulus. For accurate performance, the task requires either a 3D percept or a subject strategy that uses 2D velocity information in a manner that is computationally equivalent to that required to solve for 3D shape (Sperling et al., 1989, 1990, see the discussion of expt 2, below).

We have shown that the only cue used for the perception of three-dimensionality in these displays is motion (Sperling et al., 1989, 1990). Further experiments determined that global optic flow is used rather than the position information for individual dots, since accuracy remains high when dot lifetimes are reduced to as little as two frames (Dosher et al., 1989b). In that paper, we concluded that the input to the KDE computation is an optic flow generated by a 1st-order motion detection mechanism, such

Atch-7

as the Reichardt detector (Reichardt, 1957). Two manipulations that perturb 1st-order motion energy mechanisms—flicker and polarity alternation—also interfered with KDE (Dosher et al., 1989b). In polarity alternation, dots change over time from black to white to black on a gray background. When compared to dots that remain white, polarity alternation was equally or slightly more detectable in a detection task, was poorer but still well above chance in a discrimination of direction of motion task (computed, presumably, using tracking of the dots or using more elaborate, 2nd-order motion detection mechanisms) but was useless for tasks requiring KDE or motion segregation. These latter two tasks require the evaluation of velocity in a number of locations simultaneously (Sperling et al., 1989). Shape identification performance in a range of conditions was shown to be monotonic with a computed index of 1st-order net directional power in the stimuli (Dosher et al., 1989b) Hence, for sparse dot stimuli, KDE depends upon a simple spatio-temporal (1st-order) Fourier analysis of multiple local areas of the stimulus

In this paper, we further examine and generalize the contributions of several types of motion detectors to the optic flow computations used by the structure-from-motion mechanism

MOTION ANALYSIS MODELS AND THE KDE

1st-order motion analysis

To motivate the stimulus conditions studied here, we begin by summarizing models of early motion detection and analysis Several recent motion detection models (van Santen & Sperling, 1984, 1985, Adelson & Bergen, 1985, Watson & Ahumada, 1985) share as a common antecedent the model proposed by Reichardt (1957) We refer to this class of models as 1st-order motion detectors. Below, 2nd-order mechanisms involving additional processing stages will be discussed. In the Reichardt detector, luminance is measured at two spatial locations A and B The measurement at position A is delayed in time, and then cross-correlated over time with the measurement at position B, resulting in a "half-detector" sensitive to motion from position A to B A second such "half-detector" sensitive to motion from B to A is set in opponency with the first, resulting in the full motion detector van Santen and Sperling (1984, 1985) have investigated this model along with extensions involving voting rules for combining outputs of many detectors to enable predictions of psychophysical experiments, resulting in their Elaborated Reichardt Detector (ERD).

An alternative way of characterizing motion detection is in the frequency domain. A motion detector can be built of several linear spatio-temporal filters. Each filter is sensitive only to energy in two of the four quadrants in spatio-temporal Fourier space (ω_x, ω_t). In other words, the filters are not *separable*. Their receptive fields are oriented in space-time, and thus they are sensitive to motion in a particular direction and at a particular scale (Adelson & Bergen, 1985; Burr, Ross & Morrone, 1986, Watson & Ahumada, 1985). The Fourier "energy" (the squared output of a quadrature pair of filters) in each of two opposing motion directions is computed, and put in opponency. This "motion energy detector", proposed by Adelson and Bergen (1985), and the ERD differ in their construction and in the signals available at the subunit level, but are indistinguishable at their outputs (Adelson & Bergen, 1985, van Santen & Sperling, 1985)

The structure-from-motion computation relies upon the measurement of image velocities at several image locations The KDE shape identification task that we use here can be solved by categorizing velocity at six spatial locations into three categories leftward, approximately zero, and rightward (Sperling et al., 1989) Thus, in order to discriminate the 53 test shapes by KDE, motion detection must be followed by at least some rudimentary local velocity calculation.

In order to signal velocity, the outputs of more than one such 1st-order motion detector must be pooled Speed may be computed by pooling only two detectors (a motion and a "static" detector, Adelson & Bergen, 1985) To signal motion direction, signals must be pooled across a variety of orientations (Watson & Ahumada, 1985) Finally, in order to solve the "aperture problem" for more complex stimuli (Burt & Sperling, 1981, Marr & Ullman, 1981), signals may be pooled over a variety of directions and perhaps scales (Heeger, 1987)

In the previous paper (Dosher et al., 1989b), shape identification performance was shown to relate directly to the quality of the signal available from 1st-order motion detection mechanisms Each stimulus consisted of a large number of dots on a gray background representing a 2D projection of dots on the surface of a smooth 3D

shape under rotary oscillation. In one condition (contrast polarity alternation), the dots were first brighter than the background ("white-on-gray"), then darker than the background ("black-on-gray"), then bright again, in successive frames. For a dense random dot field (50% black/50% white) under simple planar motion, polarity alternation causes a percept of motion opposite to the true direction of motion (the "reverse-phi phenomenon", Anstis & Rogers, 1975), reverse-phi is thought to reflect a spatiotemporal Fourier analysis of the stimulus, since contrast reversal reverses the direction of motion of the lowest-frequency Fourier components (van Santen & Sperling, 1984). With contrast reversal, the outputs of 1st-order motion detection mechanisms no longer simply signal the intended direction and velocity of motion. Contrast reversal stimuli do not yield a depth-from-motion percept (Dosher et al., 1989b). We take this as evidence that the KDE relies upon input from a 1st-order motion analysis.

2nd-order motion analysis

For the sparse random dot stimuli (Dosher et al., 1989b), contrast polarity alternation eliminated the perception of structure from motion. Nonetheless, subjects could judge the direction of patches of contrast polarity alternating dots undergoing simple translation. What kind of a motion detector might be used to correctly judge the motion of a translating, polarity-alternating dot? One simple possibility would be to first apply a luminance nonlinearity to the input stimulus. For example, if the input stimulus were full-wave rectified about the mean luminance, the polarity-alternating stimulus would be converted to the equivalent of rigid motion of a white dot on a gray background. Thus, a full-wave rectifier of contrast followed by a 1st-order analyzer (such as those discussed above) would be capable of analyzing such a motion stimulus correctly (Chubb & Sperling, 1988b, 1989a, b).

A motion detection system consisting of a contrast nonlinearity followed by a 1st-order detector is one example of a wide class of "2nd-order detection mechanisms", each of which consists of a linear filtering of the input (spatial and/or temporal), followed by a contrast nonlinearity, followed by a standard 1st-order motion detection mechanism. A number of results demonstrate the existence of both 1st- and 2nd-order motion mechanisms and show the contribution of both to the perception of planar motion (Anstis & Rogers, 1975, Chubb & Sperling, 1988b, 1989a, b; Lelkens & Koenderink, 1984; Ramachandran, Rao & Vidyasagar, 1973; Sperling, 1976).

Can both 1st- and 2nd-order motion mechanisms be used by the KDE system? The polarity-alternating dots did not yield an effective KDE percept of our 3D shapes. If one accepts the existence of both 1st- and 2nd-order motion mechanisms, why didn't the 2nd-order system support KDE? The KDE stimuli were relatively small (3.7 × 4.2 deg) and viewed foveally (eye movements were permitted throughout the 2 sec stimulus duration). Evidence from studies of planar motion suggests that both systems were available under these conditions (Chubb & Sperling, 1988b). For polarity alternation stimuli, the most salient low frequency components from the 1st-order system were in the wrong direction. We assume that the 2nd-order system yields a correct (if attenuated) analysis. Bad shape identification performance may have resulted either from the perturbed 1st-order analysis or because of competition between the 1st- and 2nd-order systems (which signaled opposite directions of motion in some frequency bands). Our evidence (Dosher et al., 1989b) demonstrated that 1st-order system input is the predominant input to KDE, but it did not exclude the possibility of input from 2nd-order motion detection mechanisms. To approach that question we consider a KDE stimulus that produces a simple 2nd-order motion analysis, but to which the 1st-order motion system is, statistically, blind.

Microbalanced motion stimuli

Chubb and Sperling (1988b) defined a class of stimuli, called *microbalanced*, among which are stimuli with the properties that we desire. In expt 1 we concentrate on two examples of microbalanced motion stimuli. These stimuli are random in the sense that any given stimulus is a realization of a random process. As proven by Chubb and Sperling (1988b), if a stimulus is microbalanced then the expected output of every 1st-order detector (ERD or motion energy detector) will be zero. Thus, Chubb and Sperling defined a class of stimuli for which a consistent motion signal requires a 2nd-order motion analysis, and showed that the 2nd-order analysis predicted observers' percepts for several examples of the class.

The polarity alternation stimulus is not microbalanced; any given frequency band does show consistent motion, with the lowest spatial frequencies signalling motion in the wrong direction. This stimulus can be transformed into a microbalanced one as follows: for each dot, choose the contrast polarity randomly and independently for every frame. Any given 1st-order detector will be just as likely to signal rightward motion as it is to signal leftward motion since it will either see the same contrast polarity across any successive pair of frames or it will see contrast polarity alternate, with equal probability. One question we examine in this paper is whether the motion signal available from 2nd-order mechanisms can be used to compute 3D structure.

We present two experiments. In the first, we examine performance on a shape identification task for a variety of KDE stimuli. Several types of stimuli provide good 1st-order motion. Others are microbalanced and hence can only be analyzed by 2nd-order mechanisms. Still others offer good 1st-order motion, but involve camouflage similar to that available in some of the microbalanced conditions. We find that 1st-order motion is used, and that input from 2nd-order mechanisms may also be used but is not as robust In a second experiment, we examine the residual shape percept from two-frame KDE stimuli in order to determine whether a single velocity field is a sufficient cue for shape identification or whether acceleration also is needed

EXPERIMENT 1. POLARITY ALTERNATION, MICROBALANCE, AND CAMOUFLAGE

In the first experiment, a shape discrimination task is used with a variety of displays First, in order to sensibly compare results to our previous work (Sperling et al , 1989, Dosher et al., 1989b), there are control conditions that are identical to those of our previous experiments (the "Motion without density cue, standard speed, standard intensity" and "Motion with polarity alternation, standard speed, standard intensity" conditions of the preceding paper). In addition to dots, randomly positioned disks and lines are also used here in order to examine the effects of the foreground token used to carry the motion The disk and line tokens are larger than the single pixel dots, and hence have more contrast energy They enable us to test whether our previous failure to find KDE with polarity alternation resulted from the low contrast energy in the stimulus. Two forms of microbalanced stimuli are used, allowing us to test KDE shape identification performance with stimuli to which 1st-order motion detectors are blind. Finally, we examine stimuli in which moving textured tokens are camouflaged by a similarly textured background.

Method

Subjects. There were three subjects in this experiment. One was an author, and the other two were graduate students naive to the purposes of this experiment. All had normal or corrected-to-normal vision. There were slight differences in the conditions for each of the three subjects. These will be pointed out below

White-on-gray dot stimuli. First, we briefly describe the stimuli that consist of bright dots moving on a gray background representing a variety of 3D shapes This description will be somewhat abbreviated, since the same stimuli have been used in previous studies and more complete descriptions are available (Sperling et al , 1989) The other stimuli used in the present study result from simple image processing transformations applied to the white-on-gray dot stimuli.

Stimuli were based upon a fixed vocabulary of simple shapes consisting of bumps and concavities on a flat ground. The 3D shapes varied in the number, position, and 2D extent of these bumps and concavities. The process of generating the stimuli is illustrated in Fig 1

The first step in creating a stimulus involves the specification of a 3D surface For a square area with sides of length s, a circle with diameter $0.9\,s$ is centered, and three fixed points, labeled 1, 2 and 3, are specified For a given shape, one of two such sets of points is used (the upward-pointing triangle or the downward-pointing triangle, labeled u and d, respectively) The shape is specified as having a depth of zero outside of the circle For each of the three identified points, the depth may be either $+0.5\,s$, 0 0, or $-0.5\,s$, which are labeled as $+$, 0, and $-$, respectively The depth values for the rest of the figure were interpolated by using a standard cubic spline to connect the three interior points with the zero depth surround Thus, there are 54 ways to designate a shape u vs d, and for each of three interior points, $+$ vs 0 vs $-$. We designate a shape by denoting the triangle used, followed by the depth designations of the three points in the order shown in Fig 1A For example, $u--0$

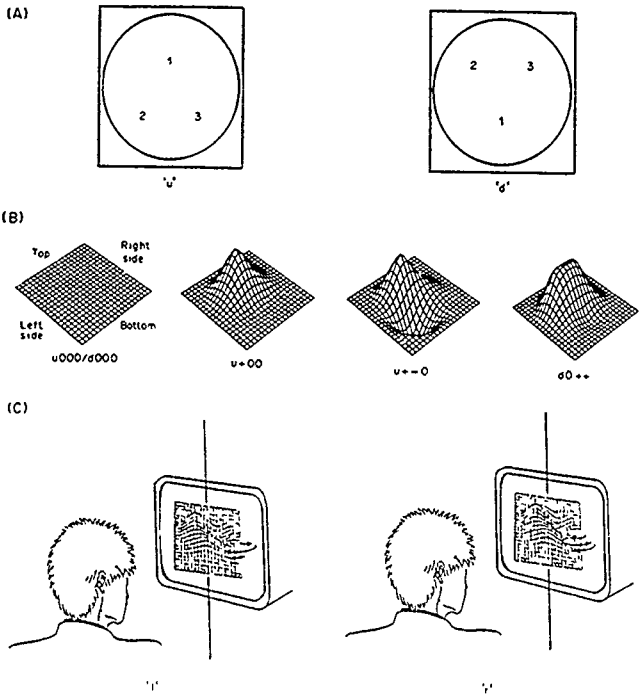


Fig 1 Stimulus shapes rotations, and their designations (A) Shapes were constructed by choosing one of the two equilateral triangles represented here Each point in the triangles was given a positive depth (i e toward the observer), zero depth, or negative depth, represented as +, 0 and −, respectively A smooth shape splined these three points to zero depth values outside of the circle A shape is designated by the choice of triangle (*u* or *d*), followed by the depth designations of the three points in the order given in the figure (B) Some representative shapes generated by this procedure All shapes consisted of a bump, concavity or both, with a variation in position and extent of these areas (C) Shapes were represented by a set of dots randomly painted on the surface of the shape, and wiggled about a vertical axis through the center of the display The motion was a sinusoidal rotation that moved the object so as to face off to the observer's right then his or her left, then back to face-forward (denoted *l*), or the reverse (denoted *r*)

is a shape with a bump in the upper-middle of the display, and a concavity in the lower-left (Fig 1B) There are 53 distinct shapes, because *u*000 and *d*000 both denote a flat square

Displays were generated by sprinkling dots randomly on the 3D surface generated by the spline, rotating that surface, and projecting the resulting dot positions onto the image plane using parallel perspective A large number of dots are chosen uniformly over a 2D area somewhat larger than the *s* by *s* square, and each dot's depth is determined by the cubic spline interpolant (where the zero depth of the surround is continued outside the square) This collection of dots is rotated about a vertical axis that is at zero depth and centered in the display. The rotation angle $\theta(k)$ is a sinusoidal "wiggle" $\theta(k) = \pm 25 \sin(2\pi k/30)$ deg, where k is the frame number within the 30 frame display Thus, the display either rotated 25 deg to the right, then reversed its direction until it faced 25 deg to the left, then reversed its direction until it was again facing forward (labeled *l*), or rotated in the opposite manner (labeled *r*, see Fig 1C) The displays presented these 3D collections of dots in parallel perspective

as luminous dots (single pixels) on a darker background

A stimulus name consists of the name of the shape followed by the type of rotation (e.g $u+-0l$), resulting in 108 possible names. Using parallel perspective, there is a fundamental ambiguity with the KDE: reversing the depth values and rotation direction of a particular shape and rotation produces exactly the same display. In other words, a convexity rotating to the right produces exactly the same set of 2D dot motions as a concavity rotating to the left. Thus, $u+-0l$ and $u-+0r$ describe precisely the same display type. There is also no difference in display type among $u000l$, $u000r$, $d000l$ and $d000r$. This results in a total of 53 distinct display types.

These experiments used 54 white-on-gray dot displays, including two instantiations of the flat stimulus $u000$ (with different dot placements) and one instantiation of each other display type. Each set of dots was windowed to a display area of 182×182 pixels (corresponding to the $s \times s$ square), with dots presented as single luminous pixels.

When the dots on the surface of a shape move back and forth in the display, the local dot density changes as the steepness of the hills and valleys changes (with respect to the line of sight). In previous work (Sperling et al., 1989), we showed that this density cue is neither necessary nor sufficient for the perception of depth. However, it is a weak cue which one of three highly trained subjects was able to use for modest above-chance performance when it was presented in isolation. In other words, changing dot density is an artifactual cue to the task. As in previous experiments, we remove this cue by deleting or adding dots as needed throughout the display in order to keep local dot density constant. As a result of this manipulation, all displays had approx. 300 dots visible in the display window. The removal of the density cue results in a small amount of dot scintillation that neither lowers performance substantially nor appears to be useful as an artifactual cue (Sperling et al., 1989, 1990).

Other tokens. The 54 stimuli described so far consisted of luminous dots moving to and fro on a less luminous background. All other stimuli were based upon these displays. First, three conditions involved changes of the token that carried the motion. The moving dots were replaced with disks, patterned disks, or wires. We refer to the dot, wire, and disk conditions as *white-on-gray* stimuli, and the patterned disks as *pattern-on-gray*.

To create a disk stimulus, a dot stimulus is modified in the following way. Each luminous dot in the stimulus is replaced with a 6×6 pixel luminous diamond centered on the dot (Fig. 2b), which appears disk-like from the viewing distance used in the experiment. A sample image of white-on-gray disks is depicted in Fig. 2c, and is based on the white-on-gray dot stimulus frame shown in Fig. 2a.

The pattern-on-gray disk stimuli are generated in a similar fashion. The 6×6 diamond consists of 24 pixels which are a mixture of black and white (12 of each). These are displayed on an intermediate gray background. The diamond pattern and a sample stimulus frame are shown in Fig. 2d and e, respectively. Note that the diamond pattern has an equal number of black and white pixels in each row.

Other stimuli were based on "wires". Each dot was connected by a straight line (subject to the pixel sampling density) to all neighbors that were at a 2D distance no greater than 15.5 pixels (Fig. 2f). Note that a vector is drawn between two points based on their distance *in the image*, not on their simulated 3D distance. Since the lines were straight, when set in motion they objectively define a thickened surface with lines cutting through the interior of each bump and concavity. This may have yielded a perceived

Fig. 2 (*opposite*). Stimulus display generation for expt. 1. (a) A single frame of a white-on-gray dots stimulus. All displays shown in this figure are based on this stimulus frame. (b) The diamond shape used to generate the disks from the dots. (c) A white-on-gray disks stimulus frame. (d) The patterned diamond for the pattern-on-gray condition. (e) A pattern on gray frame. (f) A white-on-gray wires frame. All pairs of dots in Fig. 2A were connected whose inter-point distance was less than 15.5 pixels. (g) A frame of dynamic-on-gray dots. In this condition each dot was painted black or white randomly and independently with probability of 0.5 for each color. (h) A frame of dynamic-on-gray disks. The same procedure as in (g) was applied to each pixel lying in each disk. (i) A frame of dynamic on-gray wires. (j) A frame of dynamic on static disks. For both dynamic-on-static conditions (disks and wires), the tokens and the background consisted of random dot noise and so the tokens cannot be discerned from a single static frame. (k) A frame of the pattern on-static condition. This frame contains 300 copies of the pattern in (d) on a static noise background. The camouflage is quite effective. (l) An enlargement of the central portion of (k) with the patterned disks emphasized.

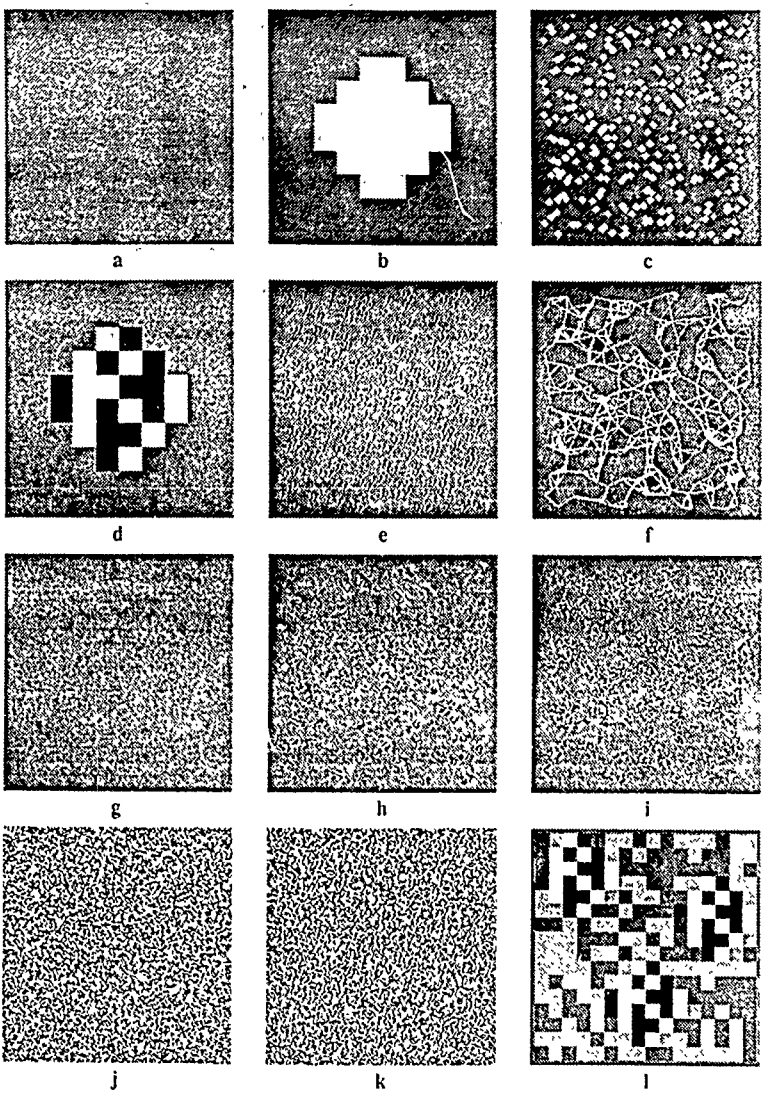


Fig 2

(tesselated) surface having slightly less relative depth than the base surface. The choice of 15.5 pixels as the criterion for drawing a line was a compromise set in order to make sure that all stimulus dots became an endpoint to at least one line, and that no line was so long as to excessively cut through the simulated surface.

The white-on-gray disks and pattern-on-gray disks were based on the dot stimuli. The same exact instantiations were used in all three conditions. The *n*th frame of a given shape and rotation consisted of either dots, disks or patterned disks centered on the same set of image positions For the wire stimuli, a new set of 54 instantiations was made.

Dynamic-on-gray. Three types of stimuli were used to explore the motion of patches of dynamic noise moving on a gray background. These stimuli are microbalanced, as we discussed in the previous section. These stimuli are derived from the dot, disk, and wire stimuli To produce a dynamic-on-gray stimulus from a white-on-gray stimulus, simply change the luminance of each white pixel in each stimulus frame (i.e the foreground or token pixels) to black randomly and independently with probability 0 5 Thus foreground pixels undergo random contrast polarity alternation while background pixels are gray (i.e have zero contrast) Sample frames are illustrated in Fig. 2g, h and i

Dynamic-on-static Two types of stimuli were used to explore the motion of patches of dynamic noise moving on a static noise background This class of stimuli is also microbalanced (Chubb & Sperling, 1988b) We derive dynamic-on-static stimuli from the disk and wire stimuli The foreground pixels consist of dynamic noise, just as in the previous dynamic-on-gray case The background pixels consist of a static frame of patterned texture, where each pixel is randomly chosen to be either black or white with a probability of 0.5, just as the dynamic noise is If a given pixel is a background position for two successive frames, then its color does not change If that position is a foreground pixel in either or both frames, then there is a 50% chance that its color will change A single frame of dynamic-on-static stimulus is simply a frame of random dot noise (Fig 2j) The motion-carrying tokens are not discernible from a single frame Rather, the areas of moving dynamic noise define the foreground tokens

Contrast polarity alteration Three stimulus conditions involved contrast polarity alternation This stimulus manipulation was explored thoroughly for dot stimuli in the preceding paper (Dosher et al., 1989b) In this condition, the motion-carrying tokens alternate from white to black to white again on successive frames, all against a background of intermediate gray. Constrast polarity alternation was used with dots, disks, and wires, resulting in three polarity alternation conditions.

Pattern-on-static. The final condition involves pattern camouflage. This condition is derived from the pattern-on-gray stimuli. The gray background is replaced with a frame of static random dot noise. In other words, the patterned disk tokens move to and fro in front of a screen of static random dots, occluding it (and occasionally each other) as they pass by. A frame of this stimulus condition is pictured in Fig 2k, and enlarged in Fig 2l, where we have artificially highlighted the patterned disks for comparison to the pattern kernel shown in Fig. 2d. There are approx. 300 patterned disks in Fig. 2k. As you can see, the camouflage is quite effective. When the patterned disks move, as one might expect, they are easily visible (Julesz, 1971).

Display details. There are a total of 13 conditions (3 white-on-gray, 1 pattern-on-gray, 3 contrast polarity alternation, 3 dynamic-on-gray, 2 dynamic-on-static, and 1 pattern-on-static) There were 54 distinct displays for each of the 13 conditions In all conditions, the displays are windowed to an area of 182 × 182 pixels Displays were computed using the HIPS image processing software (Landy, Cohen & Sperling, 1984a, b), and displayed by an Adage RDS-3000 image display system.

Subjects MSL and JBL viewed these stimuli on a Conrac 7211C19 RGB color monitor Only the green gun was used, and so stimuli appeared as bright green and black pixels (as dots, disks, lines or noise) on a green background of intermediate luminance. The stimuli subtended 3 7 × 4 2 deg Stimuli were viewed monocularly through a dark viewing tunnel, using a circular aperture which was slightly larger than the stimuli

Subject LJJ viewed the stimuli on a US Pixel PX15 black and white monitor with a P4-like phosphor Here, stimuli subtended 2 9 × 2 9 deg, and appeared as white and black pixels on an intermediate gray background Stimuli were viewed monocularly through a circular aperture in cardboard which approximately matched the hue of the displays, and

which had approximately the same luminance as the stimulus background.

Each stimulus consisted of 30 stimulus frames. These were presented at a 60 Hz frame rate. Each frame was repeated four times, resulting in an effective rate of 15 new stimulus frames per second. Each stimulus lasted 2 sec. A trial sequence consisted of a fixation spot, a blank interval, the 30 frame stimulus, and a blank. The fixation and blank lasted either for 1 sec each (subjects MSL and JBL), or 0.5 sec each (subject LJJ). The background luminance remained constant throughout the trial sequence. Subjects were free to use eye movements to actively explore the display. Stimuli were viewed from a distance of 1.6 m. After each stimulus display, subjects responded with the name of the shape and rotation direction using either a computer keyboard or response buttons

Slightly different image luminances were used for each subject. The background luminance for subjects MSL, JBL and LJJ were 31.0, 40.0 and 45.0 cd/m^2 respectively Since isolated luminous pixels were used, the appropriate unit of measurement is *extra* μcd/pixel for bright pixels, and *removed* μcd/pixel for dark pixels, all at a specified viewing distance (Sperling, 1971) Stimuli were calibrated so that extra μcd/pixel and removed μcd/pixel were equal. For subjects MSL, JBL and LJJ, these were 13.2, 19.2 and 15.7 μcd/pixel, respectively, at a viewing distance of 1.6 m Contrasts were nominally 100%

Procedure There were 13 stimulus conditions. For each condition, there were 54 stimuli (two instantiations of the flat stimulus *u*000, and one instantiation of each of the 52 other possible distinct shape, rotation combinations) This resulted in 702 stimuli, each of which was viewed once by each subject. These 702 trials were viewed in random order in six blocks of 117 trials On a given trial, a stimulus was shown, subjects keyed in their responses, and then feedback was provided so that we measured the best performance of which the subject was capable. Each block lasted approx 1 hr. Subjects ran several practice sessions on the white-on-gray dots condition before data were collected Given the mix of stimuli in a given condition, guessing base rates for the identification of shape and rotation direction were between 1/53 (for a strategy of random guessing) and 2/54 (for a strategy of always answering *u*000*l* or one of its equivalents)

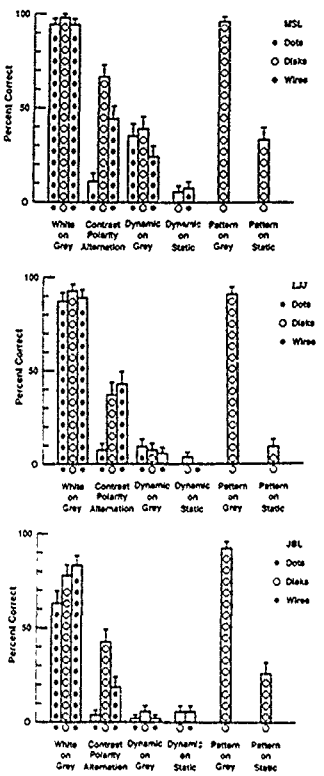


Fig 3 Results of expt 1 Results are given for three subjects Different symbols in the bars represent different tokens (large open dots for the disk and patterned disk tokens, small solid dots for the dot tokens, and asterisks for the wire tokens)

Results

The results for the three subjects are summarized in Fig 3 Each performance measure given here is the percent correct over 54 trials We discuss each class of stimulus condition in turn

White-on-gray / Pattern-on-gray. As expected, the performance on the three white-on-gray and the one pattern-on-gray condition was uniformly high The tokens provided excellent motion signals because they were moving rigid areas of high contrast It did not particularly matter whether we used dots, as in our previous studies, wires, as in the early wire-frame KDE

work (Wallach & O'Connell, 1953), disks, or patterned disks. The disk and patterned disk stimuli provided very strong percepts of shape, although the disks did not undergo realistic foreshortening as they rotated. In fact, the dot stimuli gave the weakest percepts of depth. These tokens had the least contrast energy (i.e. were the smallest), and hence were harder to detect. Subject JBL had the greatest difficulty in seeing these small dots, and his results show a slight drop in performance for the dot stimuli.

Dynamic-on-gray. The motion of a toi filled with dynamic random dot noise moving on a gray background is microbalanced. In other words, 1st-order motion detectors are "blind" to this stimulus. The expected value of the output of such a detector is zero (across random realizations of the stimulus). Simple 2nd-order mechanisms (e.g. using rectification) serve to reveal the true motion.

The results for three subjects are somewhat different. For two subjects (LJJ and JBL), performance is always at or near chance (less than 10% correct in all cases), although for subject LJJ with the dynamic-on-gray dots the performance is significantly above chance ($P < 0.05$). On the other hand, for subject MSL, performance is always well above chance (24–39% correct identifications), but far less than his nearly perfect (94–98% correct) performance with white or pattern tokens on gray.*

The 1st-order motion mechanisms are clearly the most effective input to the KDE system, since eliminating motion detectable by 1st-order mechanisms reduces performance substantially for all subjects. The results for subject MSL suggest that 2nd-order motion mechanisms can also be used. On some trials, fragments of the microbalanced stimuli did appear 3D to this subject (one of the authors), especially in the foveally-viewed portion of the stimulus. To raise his performance level, he used sophisticated guessing strategies based on active eye movements and local measurements of motion or three-dimensionality in the fov_ at a small number of locations of the display. But, these strategies only serve to bring performance up to mediocre levels in comparison with performance with rigid white-on-gray motion.

Dynamic-on-static. The dynamic-on-static manipulation also results in a micro-balanced stimulus. For the dynamic-on-static conditions, performance is at chance level for all three subjects, and for both wire disk tokens. As with the dynamic-on-gray conditions, the motion of the tokens is visible. It is not particularly difficult to detect the motion of an area of dynamic noise on a static noise background (Chubb & Sperling, 1988b). However, this sort of motion engenders no shape percept whatever under the conditions of our experiments.

Unlike dynamic-on-gray stimuli, dynamic-on-static stimuli are not revealed by contrast rectification. Detection of the motion of a region of flicker requires more elaborate 2nd-order mechanisms. Regions of flicker could first be detected by applying a linear temporal filter (such as differentiation), followed by rectification, and then by application of a 1st-order motion mechanism. Some such complex 2nd-order motion detector exists in the human visual system, since we are capable of seeing areas of flicker move, including in the displays of our experiment (at least with scrutiny). Yet, this 2nd-order motion detection system does not support the structure-from-motion computation for our dynamic-on-static stimuli.

Prazdny (1986) reached the opposite conclusion using dynamic-on-static displays representing simple wire objects rotating in a tumbling motion. Each object contained five wires, and subjects were required to identify the object among six alternative wire-frame objects.

*In order to test the range of luminances over which polarity alteration was effective, we ran a control experiment (using MSL and JBL as subjects) where a variety of white pixel luminances were used with a given black pixel luminance. We viewed a variety of dynamic-on-gray displays varying the luminance values for the black and white pixels independently over a wide range. We also tested a variety of other luminance calibration procedures. Dynamic-on-gray stimuli are only micro-balanced if the contrast energy of the white pixels is the same as that of the black pixels. And, it is difficult to calibrate the luminance of individual pixels embedded in a complex display texture given that the desired pattern is first low-pass filtered by the CRT video amplifier, and then passes through the gun nonlinearity (see Mulligan & Stone, 1989, for a full discussion of this point). Thus, it was important to verify that our results were robust over a range of luminance values overlapping the calibrated equal contrast point.

To summarize, shape identification performance is consistent with the results of expt 1 for a reasonably wide range of white pixel luminances. Subject MSL consistently performs at moderate levels, and subject JBL consistently performs at or near chance. The luminance levels yielding poor shape identification performance are consistent with the levels that result in the weakest 3D percept and are roughly consistent with the luminance levels that are balanced (black pixel decrement vs white pixel increment) for a variety of calibration displays. The performance levels for dynamic-on-gray stimuli in expt 1 do not result from a miscalibration of luminance levels.

The displays were 7 × 7 deg, and the wires were several pixels thick. Performance was quite high in the task for five subjects. Although we have some reservations about the experimental method employed by Prazdny, we have generated similar displays in our laboratory, and our dynamic-on-static wire-frame displays do yield a shape percept when displays are restricted to a small number of wires.

The most likely explanation of the difference between our results and those of Prazdny involves the difference in spatial resolution required by each task. Chubb and Sperling (1988a) have demonstrated that 2nd-order motion systems have less spatial resolution than the 1st-order mechanisms, and that their resolution drops precipitously with increases in retinal eccentricity. In our displays, motion was about a vertical axis using parallel perspective, and hence all motion was along the horizontal. There could be as many as 10 or 20 disks or wires in a given row of the image to resolve. Our displays did not yield a global percept of optic flow, but motion was perceived foveally with scrutiny. This is entirely consistent with Chubb and Sperling's observation. Prazdny did not give precise details about his stimuli, but it was clear that along a given motion path there were only two or three wires to resolve across his far larger display. Performance was so low in our dynamic-on-static conditions because too much spatial acuity was required of the 2nd-order system that detects the motion of flickering regions.

How useful for perception of shape is a display of dynamic noise figures moving on a static noise background? We have examined a large number of disk and (thick) wire displays in order to span the gap of spatial resolution between Prazdny's displays and our own. With our 3 × 3 deg display size, a shape percept can only be achieved by using a very small number of tokens (around 5–10). These displays consisted of rotating disk tokens. Cavanagh and Ramachandran (1988) suggest an alternative explanation of the difference between our results and those of Prazdny. They consider the crucial difference to be that the objects portrayed in the Prazdny displays were connected (one long wire figure), whereas our displays consisted of separate disk tokens. With our wire displays, almost no 3D percept was achieved for the dynamic-on-static condition. In addition, we were able to achieve a 3D percept with displays of a small number of dynamic-on-static disks. Thus, we feel that low spatial resolution in the 2nd-order motion system (rather than unconnected tokens) is the likely explanation for failure of KDE.

Contrast polarity alternation. Performance is quite poor for the contrast polarity-alternating dots as it was in the previous paper (Dosher et al., 1989b). For two subjects (JBL and LJJ) performance is at chance or insignificantly above chance. For subject MSL, performance is low (11% correct) but significantly above chance ($P < 0.05$). On the other hand, when the token is changed to disks or wires, performance rises substantially. Contrast polarity alternation is not as devastating a stimulus manipulation for disks and wires as it is for dots.

For 1st-order motion detection mechanisms such as the Reichardt detector, contrast polarity alternation causes the strongest responses to be in the wrong direction. Yet, the intended motion can be detected quite accurately if a 2nd-order detector is used that first applies a luminance nonlinearity followed by a Reichardt detector. The primary difference between the dots on the one hand, and the disks and wires on the other, is that the disks and wires have more pixels illuminated. In other words, they have more contrast energy, and in particular thay have more energy at lower spatial frequencies. Thus, the disk and wire stimuli should stimulate both the 1st- and 2nd-order motion detection systems more strongly, resulting in stronger incorrect direction information from the 1st-order system as a whole, but also stronger information from the 2nd-order system, and stronger directional information in those selected 1st-order frequency bands which signal the correct direction.

It is interesting to note that a large number of the errors made by observers with polarity-alternating stimuli were errors in the direction of rotation *only*, with the shape specified correctly. For example, for a stimulus which had as correct answers either $u- -0l$ or $u-+0r$, the subject incorrectly responded with $u+-0r$ or $u-+0l$, rather than with any of the 104 other possible incorrect responses. This effect was largest for the disk tokens. In a separate control experiment, for contrast polarity-alternating disk stimuli, 39% of the errors made by subject MSL were only an error in the specification of direction, compared to 1.4% direction errors for the dynamic-on-gray conditions. For subject JBL, the corresponding values were 48% and 5.6%. For the polarity-alternating disks, on

trials when subject MSL correctly identified the shape, there was a 33% chance that he would misidentify the direction of rotation (for JBL: 29.3%). We believe that accurate shape identification in this condition primarily reflects responses constructed from selected 1st-order information. One strategy was simply to specify the opposite rotation direction to that which was perceived! The displays did, however, occasionally appear to be 3D with the correct direction of motion (at certain times during the rotation, or close to the location to which the eyes were directed), indicating a residual 2nd-order motion input to the KDE system. The fact that these displays only appeared foveally to be rotating in the correct direction, and then only using the larger tokens, is consistent with a 2nd-order motion detection system with low contrast sensitivity and low spatial resolution (as has been demonstrated by Chubb & Sperling, 1988b), and more sensitive in the fovea (Chubb & Sperling, 1988a). In summary, we have some indication that 2nd-order motion detection mechanisms can be used to derive 3D structure, but they are far less robust and have poorer spatial resolution than 1st-order motion mechanisms.

Pattern-on-static. For all three subjects performance with pattern-on-static displays is quite poor (9, 26 and 33% correct), although it is significantly above chance levels in all cases ($P < 0.05$). This poor performance results from a mismatch of resolution and temporal sampling. The patterned disks are quite detailed/high frequency. The disks are 6 pixels in diameter, and can move as far as 8.3 pixels in one frame. This speed is only achieved by disks at the top of a peak when in the middle of the display (i.e. near frame numbers 0, 15 and 29), but many disks are moving 3–5 pixels per frame. High frequency spatial filters which are required to identify the disks must correlate across frames with filters that are far more than 90 deg away in the phase of their peak spatial frequency. A typical 1st-order detector will not compare spatial regions that far apart in order to avoid spatio-temporal aliasing (van Santen & Sperling, 1984). Thus, the clearest motion signals are coming from the slower areas in the display, which are the least useful for discriminating the shapes. We have examined pattern-on-static displays with finer temporal sampling (60 new frames per sec, as opposed to 4 repaints of 15 new frames per sec used in the experiment), and they give a strong impression of three-dimensionality. Thus, poor performance in the task resulted from undersampling in time of the stimuli, which interferes with 1st-order (and some 2nd-order) motion mechanisms, and good KDE can result from the motion of tokens which are camouflaged when at rest.

We have also examined dynamic-on-static displays with finer temporal sampling (60 new frames per sec). These displays yield no impression of three-dimensionality. The poor results for dynamic-on-static displays do not result from insufficient sampling in time. Also, since finely sampled pattern-on-static displays do appear 3D, poor performance with dynamic-on-static-displays does not result from the camouflage of the tokens when at rest. Rather, dynamic-on-static displays yield no effective KDE because of the low resolution of the 2nd-order system required to analyze the motion.

EXPERIMENT 2: TWO-FRAME KDE

The first experiment shows that accurate performance in shape identification is dependent upon a global (primarily 1st-order) optic flow. If a stimulus manipulation makes that optic flow noisy or otherwise interferes with the optic flow computation, there is little or no KDE. This occurs even though foveal scrutiny does reveal the motion in these displays.

If the percept of surface shape depends upon a global optic flow, then we should be able to get reasonable shape identification performance from any stimulus that results in a strong percept of optic flow. In particular, the extended (2 sec) viewing conditions of expt 1 should not be necessary. Two frames are obviously the minimum number of frames that can yield a percept of motion, and two frames should suffice. In the second experiment, we investigate the accuracy of performance in the shape identification task for two-frame displays.

Method

Subjects. There were two subjects in this experiment. One was an author, and the other was a graduate student naive to the purposes of this experiment. Both had normal or corrected-to-normal vision. There were slight differences in the conditions for each of the two subjects. These will be pointed out below.

Stimuli and apparatus. The stimuli were similar to the white-on-gray dot stimuli from expt 1. Stimuli were generated from the same set of 3D

shapes, using the same dot densities, and projected in the same way. The local dot density was kept constant using the same scintillation procedure. New stimuli were computed, two of the flat shape, and one of each of the other 52 shapes, resulting in 54 displays.

Each display consisted of 11 frames, rotating from 20 deg left to 20 deg right in increments of 4 deg per frame. The middle frame (number 6) was face-forward, as was the first frame of each display in expt 1. Two-frame stimuli consisted of a presentation of the middle frame followed by one of the other 10 display frames. This resulted in either a leftward or rightward rotation of 4–20 deg between the two frames of the display. A single trial display consisted of 0.5 sec of a cue spot, 0.5 sec blank, the first frame, an inter-stimulus blank interval (or ISI), the second frame, and a blank. Each stimulus frame was repainted four times at 60 Hz, for a total duration of 67 msec. We define the ISI to be the time interval between the onset of the last painting of the first stimulus frame and the onset of the first painting of the second stimulus frame. For example, when no blank frames were used, the ISI was 16.7 msec. Displays were 182 × 182 pixels, and were presented using the same apparatus and viewing conditions as for subject LJJ in expt 1. The background luminances for subjects MSL and LJJ were 15.6 cd/m^2 and 5.0 cd/m^2, respectively. The corresponding dot luminosities were 26.8 and 15.7 extra μcd/dot, respectively. Nominal contrasts were huge (i.e. nominal Weber contrasts of 500% or more).

Procedure. The task was shape and rotation identification. Subjects keyed their responses using response buttons, and received feedback on the display after their response. Three groups of trials were run. In the first, the ISI was 16.7 msec, and rotation angle between frames was varied from 4 to 20 deg. Since the second frame could be chosen from either the frames preceding or succeeding the middle frame (rotation to the left or right), this resulted in 540 possible stimuli (54 displays, 2 directions, 5 rotation angles). These were run in random order in 4 blocks of 135 trials. In the second group of trials, rotation was kept constant at 4 deg. ISI ranged from 16.7 to 83.3 msec. This again resulted in 540 trials presented in random order in 4 blocks of 135 trials. In the third group

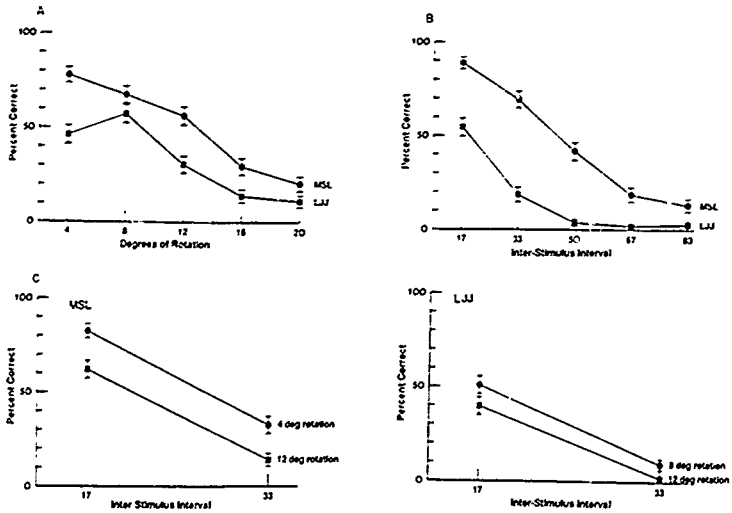


Fig 4. Results of expt 2 Data for two subjects are shown Error bars indicate ±1 SEM (A) Shape-and-rotation identification accuracy as a function of the angle of rotation between the two frames ISI was 16 7 msec (B) Shape-and-rotation identification accuracy as a function of the duration of a blank inter-stimulus interval (ISI) Rotation angle was 4 deg (C) The two manipulations used in the same experiment Note the lack of interaction

of trials, both rotation angle and ISI were varied. The ISIs were either 16.7 or 33.3 msec. For subject MSL, the rotation angles were either 4 or 12 deg. For LJJ, they were either 8 or 12 deg. These four conditions (two rotation angles by two ISIs) resulted in 432 trials which were presented in random order in 4 blocks of 108 trials.

Results

The results are shown in Fig. 4. Each data point is the percent correct over 108 trials. As is evident from the figure, shape identification can be quite high for these minimal motion displays (for similar observations using different experimental methodology, see Braunstein, Hoffman, Shapiro, Andersen & Bennett, 1987, Lappin, Doner & Kottas, 1980, Mather, 1989, and Petersik, 1980). For an ISI of 16.7 msec (Fig. 4A), this entire sequence lasted only 133 msec. Yet, performance was as high as 54.6% for subject LJJ, and 88.9% for subject MSL (62.8% and 94.2% of their white-on-gray dots performance in expt 1, respectively). Two frames of moving dots are sufficient for accurate, although not perfect
performance in this shape identification task. Since these experiments were first reported (Landy, Sperling, Dosher & Perkins, 1987a, Landy, Sperling, Perkins & Dosher, 1987b), Todd (1988) has also shown above-chance KDE performance for two-frame stimuli, although in his paradigm the two frames are repeated several times before a response is made.

Rotation angle and fixation. Performance as a function of rotation angle between the two frames is given in Fig. 4A. Performance decreases with increasing angle of rotation for subject MSL. For subject LJJ, performance reaches a peak at 8 deg, and decreases for smaller and larger rotations. The decrease in performance with larger rotation angles is to be expected, since the correspondence problem becomes increasingly difficult as dots move farther from their initial positions. One might also expect performance to drop as rotation angle decreases to zero. At extremely small rotation angles, the remaining motion would fall below threshold. In our displays, the drop with small rotation angles might be expected to occur even sooner as the small motions in the display became corrupted by poor spatial sampling (inter-pixel distance was approx. 1 min arc). This drop was only seen in the data of LJJ, and presumably would be seen in those of MSL if he had been tested using smaller rotations.

In a previous paper (Dosher et al., 1989b), we found that adding a blank interval between successive frames of a 30 frame KDE stimulus reduced shape identification to near chance performance. This was explained by reduction of power in the stimulus to the 1st-order system. This effect is also seen here, where performance decreases monotonically with increasing ISI (Fig. 4B). Subject LJJ performs at chance levels with a 50 msec or greater ISI, while subject MSL is still slightly above chance performance with an 83.3 msec ISI.

Time and distance. In the previous two groups of trials, there was a confounding between the stimulus manipulation (rotation angle or ISI) and dot velocity. Greater rotation angles at a fixed (16.7 msec) ISI produced greater velocities. Similarly, greater ISIs at a fixed 4 deg rotation angle resulted in smaller velocities. If performance were simply a function of velocity, then rotation angle and ISI should trade off. In Fig. 4C we present the results of varying both ISI and rotation angle factorially. We used a different set of rotations for subject LJJ than MSL based on the results in Fig. 4A, so that for both subjects the performance was expected to decrease with increasing rotation angles. As can be seen in the figure, the two variables do not trade off as would be expected if performance were only a function of velocity, or rotation speed. Increasing rotation angle increases the difficulty of the correspondence problem. Increasing ISI causes increasing problems for the motion detection system. Both manipulations degrade performance in an additive fashion. This observation contradicts Korte's (1915) 3rd law of apparent motion perception, which states that an increase in ISI must be counteracted by an increase in distance traveled for strong apparent motion. In Fig. 4C, Korte's law predicts a cross-over interaction, which is strongly disconfirmed. However, Burt and Sperling (1981) show that time and distance have independent additive effects on the strength of the apparent motion of dot stimuli, which agrees with the present results.

KDE from optic flow. Accurate KDE performance requires a global optic flow. When that optic flow is produced by a minimal motion stimulus—a two-frame display—the shape percept may be fragile and easily degraded by a variety of stimulus manipulations. The stimuli are quite brief in this paradigm and, by subject

reports, appear as a collection of dots moving at various speeds, i.e. "look like" an optic flow. On some trials, only patches of planar motion are perceived, and the shape response is generated cognitively. On other trials, a 3D surface is perceived. On some trials the optic flow is perceived and so is the shape, but the shape percept is only "felt" after the display is over. As we discussed extensively in our first article on the shape identification task (Sperling et al., 1989), KDE is inextricably tied with the percept of an optic flow. It can be very difficult to differentiate empirically between a judgment based on a 3D percept and performance based on an alternative strategy (computationally equivalent to that required for KDE) using a remembered set of 2D velocities.

Reasonably accurate performance on the shape-and-rotation identification task results from only two frames of 300 points. In the computer vision literature, there have been several studies of the structure-from-motion problem resulting in theorems of the following form "*m* views of *n* points under the following restrictions of the motion path suffice to determine the 3D structure up to a reflection" (Bennett & Hoffman, 1985, Hoffman & Bennett, 1985, Hoffman & Flinchbaugh, 1982, Ullman, 1979) It has been suggested that these minimal conditions for structure from motion also govern human perception (Braunstein et al., 1987, Petersik, 1987) The particular models just mentioned do not have any prediction concerning performance in the 300 points 2 views situation used here An exception is a recent paper by Bennett, Hoffman, Nicola and Prakash (1989), where it is shown that there is a one parameter family of possible interpretations for two frames of four or more points This family is parameterized by the slant of the axis of rotation (as in the "isokinescopic displays" described by Adelson, 1985), and the paper does not deal explicitly with rotation axes in the image plane, as used here. On the other hand, models that compute 3D structure based only upon a single velocity field do allow for this performance (Longuet-Higgins & Prazdny, 1980, Koenderink & van Doorn, 1986) We take our experimental results as evidence for optic flow-based methods for the KDE, as opposed to models requiring three or more views In particular, our results strongly rule out models that require measurement of acceleration in addition to velocity (e.g Hoffman, 1982)

Structure-from-motion computation may improve its 3D representation with additional information (e.g. with additional frames, Grzywacz, Hildreth, Inada & Adelson, 1988; Hildreth & Grzywacz, 1986; Landy, 1987; Ullman, 1984). The shape in our two-frame displays does not always appear to have the depth extent that results from the 30 frame displays of expt 1, and two-frame performance is reduced relative to 30-frame performance. The shape identification task can be solved by knowing only the sign of depth and direction of motion in each spatial location (up to a reflection), without accurately estimating either velocity or the amount of depth.

DISCUSSION

Two experiments investigated the type of motion detection mechanism used as an input to the structure-from-motion system. Performance in the shape-and-rotation identification task was accurate regardless of the token used to carry the motion, as long as that token was presented with constant contrast polarity (the white-on-gray and pattern-on-gray conditions). The performance decrements seen with contrast polarity alternation and the two microbalanced conditions add further evidence to the conclusion of Dosher et al (1989b) that 1st-order motion detectors are the primary substrate for the computation of shape. In addition, there are indications of an input to the shape computation from 2nd-order motion mechanisms, which is weak, low in spatial resolution, and concentrated at the fovea. 2nd-order mechanisms that require temporal filtering (i.e detection of flicker) prior to a point nonlinearity were useless here because of the spatial resolution required by our stimuli. These sorts of detectors would only be useful for KDE displays involving a small number of moving features, rather than the densely sampled optic flows required for the determination of precise shapes of curved surfaces from motion cues The results from the two-frame experiments reinforced these conclusions. They also demonstrated that detection of instantaneous velocity is sufficient for KDE, acceleration is not required, nor are more than two views.

Acknowledgements—The work described in this paper was supported primarily by a grant from the Office of Naval Research grant N00014-85-K-0077, and partly by USAF Life Science Directorate, grants 85-0364, 88-0140, and NSF grant IST-8418867 We would like to thank Charles Chubb

for his helpful comments, and Robert Picardi for technical assistance. Portions of this work have been presented at the annual meetings of the Association for Research on Vision and Ophthalmology, Sarasota, Florida (Landy et al., 1987a) and the Optical Society of America, Rochester, New York (Landy et al., 1987b).

REFERENCES

Adelson, E. H. (1985) Rigid objects appear highly nonrigid. *Investigative Ophthalmology and Visual Science* (Suppl.), *26*, 56.

Adelson, E. H. & Bergen, J. R. (1985) Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, *2*, 284–299.

Anstis, S. M. & Rogers, B. J. (1975) Illusory reversal of depth and movement during changes of contrast. *Vision Research*, *15*, 957–961.

Bennett, b. M. & Hoffman, D. D. (1985) The computation of structure from fixed-axis motion: Nonrigid structures. *Biological Cybernetics*, *51*, 293–300.

Bennett, B. M., Hoffman, D. D., Nicola, J. E. & Prakash, C. (1989) Structure from two orthographic views of rigid motion. *Journal of the Optical Society of America A*, *6*, 1052–1069.

Braunstein, M. L., Hoffman, D. D., Shapiro, L. R., Andersen, G. J. & Bennett, B. M. (1987) Minimum points and views for the recovery of three-dimensional structure. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 335–343.

Burr, D. C., Ross, J. & Morrone, M. C. (1986) Seeing objects in motion. *Proceedings of the Royal Society of London B*, *227*, 249–265.

Burt, P. & Sperling, G. (1981) Time, distance and feature trade-offs in visual apparent motion. *Psychological Review*, *88*, 171–195.

Cavanagh, P. & Ramachandran, V. S. (1988) Structure from motion with equiluminous stimuli. Paper presented to the *Annual Meeting of the Canadian Psychological Association*, Montreal, June.

Chubb, C. & Sperling, G. (1988a) Processing stages in non-Fourier motion perception. *Investigative Ophthalmology and Visual Science* (*Suppl.*), *29*, 266.

Chubb, C. & Sperling, G. (1988b) Drift-balanced random stimuli: A general basis for studying non-Fourier motion perception. *Journal of the Optical Society of America A*, *5*, 1986–2007.

Chubb, C. & Sperling, G. (1989a) Two motion perception mechanisms revealed through distance-driven reversal of apparent motion. *Proceedings of the National Academy of Sciences U.S.A.*, *86*, 2985–2989.

Chubb, C. & Sperling, G. (1989b) Second order motion perception: Space/time separable mechanisms. *Proceedings Workshop on Visual Motion* (pp. 126–138). Washington, D.C.: IEEE Computer Society Press.

Dosher, B. A., Landy, M. S. & Sperling, G. (1989a) Ratings of kinetic depth in multi-dot displays. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 816–825.

Dosher, B. A., Landy, M. S. & Sperling, G. (1989b) The kinetic depth effect and optic flow—I. 3D shape from Fourier motion. *Vision Research*, *29*, 1789–1813.

Grzywacz, N. M., Hildreth, E. C., Inada, V. K. & Adelson, E. H. (1988) The temporal integration of 3-D structure from motion. A computational and psychophysical study. In von Seelen, W., Shaw, G. & Leinhos, U. M. (Eds.), *Organization of neural networks*. New York: VCH.

Heeger, G. J. (1987) Model for the extraction of image flow. *Journal of the Optical Society of America A*, *4*, 1455–1471.

Hildreth, E. C. & Grzywacz, N. M. (1986) The incremental recovery of structure from motion: Position vs. velocity based formulations. *Proceedings of the workshop on motion: Representation and analysis*. IEEE Computer Society no. 696, Charleston, South Carolina, 7–9 May.

Hoffman, D. D. (1982). Inferring local surface orientation from motion fields. *Journal of the Optical Society of America*, *72*, 888–892.

Hoffman, D. D. & Bennett, B. M. (1985) Inferring the relative three-dimensional positions of two moving points. *Journal of the Optical Society of America A*, *2*, 350–353.

Hoffman, D. D. & Flinchbaugh, B. E. (1982) The interpretation of biological motion. *Biological Cybernetics*, *42*, 195–204.

Julesz, B. (1971) *Foundations of cyclopean perception*. Chicago, IL: The University of Chicago Press.

Koenderink, J. J. & van Doorn, A. J. (1986) Depth and shape from differential perspective in the presence of bending deformations. *Journal of the Optical Society of America A*, *3*, 242–249.

Korte, A. (1915) Kinematoskopische Untersuchungen. *Zeitschrift für Psychologie*, *72*, 193–206.

Landy, M. S. (1987) A parallel model of the kinetic depth effect using local computations. *Journal of the Optical Society of America A*, *4*, 864–876.

Landy, M. S., Cohen, Y. & Sperling, G. (1984a) HIPS: A Unix-based image processing system. *Computer Vision, Graphics and Image Processing*, *25*, 331–347.

Landy, M. S., Cohen, Y. & Sperling, G. (1984b) HIPS: Image processing under UNIX—Software and applications. *Behavior Research Methods, Instruments and Computers*, *16*, 199–216.

Landy, M. S., Sperling, G., Dosher, B. A. & Perkins, M. E. (1987a) Structure from what kinds of motion? *Investigative Ophthalmology and Visual Science* (*Suppl.*), *28*, 233.

Landy, M. S., Sperling, G., Perkins, M. E. & Dosher, B. A. (1987b) Perception of complex shape from optic flow. *Journal of the Optical Society of America A*, *4*, 106.

Lappin, J. S., Doner, J. F. & Kottas, B. L. (1980) Minimal conditions for the visual detection of structure and motion in three dimensions. *Science*, *209*, 717–719.

Lelkens, A. M. M. & Koenderink, J. J. (1984) Illusory motion in visual display. *Vision Research*, *24*, 1083–1090.

Longuet-Higgins, H. C. & Prazdny, K. (1980) The interpretation of a moving retinal image. *Proceedings of the Royal Society of London B*, *208*, 385–397.

Marr, D. & Ullman, S. (1981) Directional selectivity and its use in early visual processing. *Proceedings of the Royal Society of London B*, *211*, 151–180.

Mather, G. (1989) Early motion processes and the kinetic depth effect. *The Quarterly Journal of Experimental Psychology*, *41A*, 183–198.

Mulligan, J. B. & Stone, L. S. (1989) Halftoning method for the generation of motion stimuli. *Journal of the Optical Society of America A*, *6*, 1217–1227.

Petersik, J. T. (1980) The effects of spatial and temporal factors on the perception of stoboscopic rotation simulations. *Perception*, *9*, 271–283.

Petersik, J. T. (1987). Recovery of structure from motion: Implications for a performance theory based on the structure-from-motion theorem. *Perception and Psychophysics, 42,* 355–364.

Prazdny, K. (1986). Three-dimensional structure from long-range apparent motion. *Perception, 15,* 619–625.

Ramachandran, V. S., Rao, V. M. & Vidyasagar, T. R. (1973). Apparent movement with subjective contours. *Vision Research, 13,* 1399–1401

Reichardt, W. (1957). Autokorrelationsauswertung als Funktionsprinzip des Zentralnervensystems. *Zeitschrift Naturforschung B, 12,* 447–457.

van Santen, J. P. H. & Sperling, G. (1984) Temporal covariance model of human motion perception. *Journal of the Optical Society of America A, 1,* 451–473

van Santen, J P. H & Sperling, G (1985) Elaborated Reichardt detectors. *Journal of the Optical Society of America A, 2,* 300–321

Sperling, G (1971) The description and luminous calibration of cathode ray oscilloscope visual displays *Behavior Research Methods and Instruments, 3,* 148–151

Sperling, G (1976) Movement perception in computer-driven visual displays. *Behavior Research Methods and Instrumentation, 8,* 144–151.

Sperling, G., Landy, M. S., Dosher, B. A & Perkins, M. E. (1989). The kinetic depth effect and identification of shape. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 826–840.

Sperling, G., Dosher, B. A. & Landy, M. S (1990). How to study the kinetic depth effect experimentally. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 445–450.

Todd, J. T. (1988). Perceived 3D structure from 2-frame apparent motion *Investigative Ophthalmology and Visual Science (Suppl.), 29,* 265.

Ullman, S. (1979). *The interpretation of visual motion* Cambridge, MA: MIT Press

Ullman, S (1984) Maximizing rigidity: The incremental recovery of 3-D structure from rigid and non-rigid motion *Perception, 13,* 255–274

Wallach, H. & O'Connell, D. N (1953). The kinetic depth effect *Journal of Experimental Psychology, 45,* 205–217.

Watson, A. B. & Ahumada, A J. Jr (1985). Model of human visual-motion sensing *Journal of the Optical Society of America A, 1,* 322–342

Anne Sutter, George Sperling, and Charles Chubb. Measuring the Spatial Frequency Selectivity of Second-Order Texture Mechanisms. Investigative Opthalmology and Visual Science, 1990, **31, No. 4,** *ARVO Supplement, 104*

MEASURING THE SPATIAL FREQUENCY SELECTIVITY OF SECOND-ORDER TEXTURE MECHANISMS

Anne Sutter, George Sperling, & Charles Chubb

Human Information Processing Laboratory, New York University, NY, NY 10003

Recent studies of texture and motion perception suggest two parallel processing systems: a first-order system consisting of selective linear filters followed immediately by detectors, and a second-order system in which preprocessing (consisting of an initial stage of linear filtering followed by rectification) precedes subsequent stages of selective linear filtering and detection. Here we measure two properties of the second-order system: the contrast modulation sensitivity as a function of spatial frequency (MTF) of its second-stage filters, and the relation of initial spatial filtering to second-stage selectivity. To determine the MTF, amplitude modulation thresholds were determined for Gabor modulations of a carrier noise. The carrier was spatially bandlimited noise with an approximate bandwidth of one octave. Four carrier bands were created with center frequencies of 2, 4, 8, and 16 c/deg. The spatial frequency of the test signals (imposed amplitude modulations) ranged from 0.5 to 8 c/deg. We used a staircase procedure that required subjects to specify the orientation (vertical or horizontal) of the modulating signal.

Results. (1) The threshold amplitude of signal modulation was lowest for 0.5 and 1.0 c/deg. Above 1.0 c/deg, threshold increased with frequency. (2) Threshold modulation was independent of the spatial frequency of the carrier noise[1]. (3) There was no significant interaction of carrier frequency band with the modulating frequency. These results indicate that the second-stage selective filters and detectors are most sensitive to frequencies less than or equal to 1 c/deg but that they are indifferent to the spatial frequency content of the carrier noise upon which these signals are impressed.

[1] Jamar, J.H.T. & Koenderink, J.J., (1985) *Vis Res* 25 (4) pp 511-521

Supported by AFOSR Life Sciences Directorate Grant 88-0140 and NIMH Grant ST32MH14267

Atch- 8

Joshua A. Solomon, Charles Chubb, and George Sperling. The Lateral Inhibition of Perceived Textural Contrast is Orientation Specific. Investigative Opthalmology and Visual Science, 1990, 31, No. 4, *ARVO Supplement, 561*

THE LATERAL INHIBITION OF PERCEIVED TEXTURAL CONTRAST IS ORIENTATION SPECIFIC

Joshua A. Solomon, Charles Chubb, and George Sperling.*
Human Information Processing Laboratory, New York University.
*Psychology Department, Rutgers University

For a test patch of isotropic spatial texture P embedded in a surrounding texture field S, the perceived contrast of P depends substantially on the contrast of the texture surround S.[1] When P is surrounded by a high contrast texture with a similar spatial frequency content, it appears to be less contrasty than when it is surrounded by a uniform field. Here we demonstrate that this lateral suppression of P's apparent contrast by the surrounding texture S is *orientation specific*. That is, suppression of apparent contrast of a patch of sinusoidal grating P by a surround grating S of the same spatial frequency is greatest when the angle between gratings P and S is 0 deg. Using dynamically phase-shifting sinusoidal gratings of 3.3, 10 and 20 c/deg, we measured orientation-specific suppression of apparent contrast at two levels of contrast *Results* (1) Both parallel and orthogonal S gratings caused suppression of P's apparent contrast relative to a uniform surround. (2) There was orientation specificity (greater contrast inhibition by 0 than 90 deg surrounds) for all $S-P$ combinations except the high-contrast 3.3 c/deg grating and the low contrast 20 c/deg grating (which was invisible). (3) Orientation specificity increased with greater spatial frequencies and with lower stimulus contrasts. The results suggest a contrast perception mechanism in which both oriented and nonoriented units determine the perceived lightness or darkenss of a point in visual space, and every unit is inhibited primarily by similar adjacent units

[1]Chubb, C., Sperling, G., & Solomon, J. A. (1989) Proc. Natl Acad. Sci USA 86, 9631-9635

Supported by AFOSR Life Sciences, Visual Information Processing Program, Grant 88-0140

Atch-9

Vision Res Vol 30, No 6 pp 921-936 1990
Printed in Great Britain

0042-6989/90 $3 00 + 0 00

THE VISIBLE PERSISTENCE OF STIMULI IN STROBOSCOPIC MOTION

JOYCE E FARRELL,* M. PAVEL† and GEORGE SPERLING
New York University, Washington Square, New York, NY 10012, U S A

(*Received* 14 *November* 1988; *in revised form* 25 *September* 1989)

Abstract—This paper reports an improved paradigm to measure visible persistence The stimulus is a pair of lines stroboscopically displayed in successive positions moving in opposite directions The subjects' judgement of simultaneous appearance of all the presented lines is used to estimate visible persistence This paradigm permitted independent manipulation of spatial and temporal stimulus separations in linear motion The resulting estimates of visible persistence increase with spatial separation up to 0 24 deg of visual angle and approaches a maximum value at larger spatial separations. The results are consistent with the existence of a hypothetical visual gain mechanism that operates over small retinal distances to effectively decrease persistence duration with decreasing spatial separation

Visible persistance Stroboscopic motion Apparent motion

INTRODUCTION

Stroboscopic motion

In artificial representations of natural object motion, such as in movies, television, and computer driven visual displays, continuous motion is represented by a succession of discrete samples By increasing the temporal sampling rate of an object moving at a fixed velocity, one can create an illusion of motion that is indistinguishable from the appearance of continuous motion (Sperling, 1976, Watson, Ahumada & Farrell, 1983) When the sampling rate is not high enough, however, the appearance of continuous motion is replaced by multiple images of the moving object

Consider, for example, the stroboscopic representation of a single vertical line moving horizontally across a display screen For some spatial and temporal separations of the line in stroboscopic motion, instead of a single line, observers perceive a number of lines moving together across the screen (Allport, 1968). An analogous phenomenon in real motion is the apparent elongation of a rapidly moving object (Newton, 1720, Allen, 1926) The obvious explanation for the apparent multiple lines in stroboscopic motion and the smearing in real motion is that each flash of the line produces an image whose visibility persists over time and which, therefore, temporally overlaps subsequent flashes of the line.

According to this explanation, the visible persistence of an image can be estimated by the number of successive stimuli that appear to be simultaneous. For example, if a stimulus is visible for approx. 100 msec, it should appear to temporally overlap stimuli that follow in less than 100 msec. Previous estimates of the duration of visible persistence based on this method range between 100 and 300 msec (Coltheart, 1980) When the distance and time between successive stimuli approaches zero, as in the case of real motion, the duration of visible persistence can be estimated by the length of an object's blur streak Estimates of the duration of visible persistence based on this latter method (Burr, 1980) range between 2 and 5 msec Apparently, the procedure for investigating the persistence of stroboscopically moving stimuli generates a different estimate of persistence duration than the procedure for investigating the persistence of continuously moving stimuli. But should we attribute this difference to differences in the paradigms used for estimating persistence duration? Or do different perceptual mechanisms underlie the visible persistence of stimuli in stroboscopic ("apparent motion") and continuous ("real") motion?

*To whom reprint requests should be addressed, present address Hewlett-Packard Laboratories, P O Box 10490, Palo Alto, CA 94303-0971, U S A
†Present address, Department of Psychology, Stanford University, Stanford CA 94305, U S A

Farrell (1984) estimated the visible persistence of stimuli in stroboscopic motion by asking observers to report the number of successively presented stimuli that appeared to be simultaneously visible. She found that the estimated durations of visible persistence increased with the distance separating the successive stimuli. This finding, taken together with reports by Dixon and Hammond (1972), Allport (1970) and DiLollo and Hogben (1985), provides an explanation for the paradox that the visible persistence of continuously moving stimuli is relatively short (Burr, 1980) when compared to the persistance of stroboscopically moving stimuli (Allport, 1970, Efron & Lee, 1971) When the distance between successive stimuli is small, the duration of visible persistence is small, as the distance increases, persistence increases This reduces the smear generated by moving objects but extends the time available to process stationary objects (e.g Burr, 1980, DiLollo, 1980, Sperling, 1967)

Because visible persistence can have many different causes, it is important to determine whether lawful behavior measured using one paradigm extends to other procedures In this paper, we first review some previous methods for estimating visible persistence. We then describe a new procedure that we believe overcomes some of the limitations of the previous procedures Using our new method, we extend the measurements made by Farrell (1984) and by DiLollo and Hogben (1985) by investigating the duration of visible persistence over a wide range of spatial separations The new data that we report in this paper sheds light on the type of mechanism that may underlie the visible persistence of moving stimuli and the range over which the mechanism operates

Paradigms for estimating the duration of visible persistence

The duration of visible persistence of an object in stroboscopic motion can be estimated by the number of successive objects that appear to be physically present at the same time (Allport, 1968; Dixon & Hammond, 1972; Efron & Lee, 1971) Here, we consider the hypothesis that for describing the appearance of stroboscopically moving objects, the visual system can be represented by two stages The first stage represents low level perceptual units and is represented by a spatio-temporal filter whose response embodies visible persistence, it lengthens the duration of its visual inputs The second stage monitors the perceptual units of the first stage and decides which of the units are active by comparing their output to a threshold The number of simultaneously active units corresponds to the number of simultaneously visible stimuli. For example, suppose that a briefly presented luminous line elicits a visual sensation (the first stage response) that decays, and, after 100 msec, the persisting sensation is no longer visible (below threshold of the second stage) Suppose also that the line is presented every 100 msec in a new position, as illustrated in Fig. 1. This system will report that it sees only one line because the visible persistence of successive stimuli does not overlap. When the line is represented every 50 msec, the system reports seeing two lines because the visible persistence of two successive stimuli will overlap. By the same reasoning, the system will report 3 lines when the line is presented every 33 msec and 4

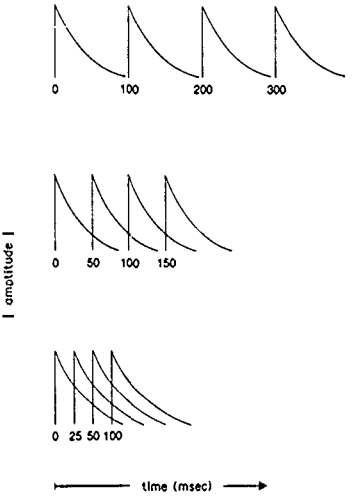


Fig 1 This figure illustrates the hypothetical case in which a briefly presented visual stimulus creates a persisting sensation that decays over time such that after 100 msec the persistence decays to a level below which it is no longer visible In the top panel, the stimulus is presented in a new position every 100 msec and a single line should appear to be present at any one instant in time The second and third panels show instances in which successively presented stimuli generate visual responses that overlap in time In general if the perceived number of stimuli increases linearly with the rate of stimulus presentation, then the slope of the linear function can be used to estimate the duration of visible persistence

lines when the line is presented every 25 msec. In general, when the perceived number of lines increases linearly with the rate of stimulus presentation, then the slope of the linear function can be used to estimate the duration of visible persistence.

Allport (1968, 1970) and Efron and Lee (1971) estimated the duration of visible persistence from the number of simultaneously visible lines by means of a computation very similar to that embodied by the 2-stage system described above. For example, Efron and Lee (1971) assumed that visible persistence can be described by a single real number, its duration p. Efron and Lee reasoned that the number of stimuli that will appear to be simultaneous is $n = p/t$ where t is the time interval separating two adjacent stimuli, and n is the average number of observed lines. Implicitly, this prediction assumes that the probability that the number of successive stimuli will appear to be simultaneously visible is proportional to the degree to which the visible persistence of successive stimuli overlap. Let the number of lines simultaneously observed on a particular trial be a random variable N and let n be the expected value of N. These assumptions lead to the prediction that:

$$n = E(N) = \max\left(\frac{p}{t}, 1\right).$$

When $p \leqslant t$, the expected value of N, $E(N)$, is 1 representing the fact that observers report seeing a stimulus even when it is not visible all the time. When $p > t$, $E(N)$ is p/t. This prediction is precisely correct only for integer values of p/t (see below).

Efron and Lee (1971) varied the rate at which a rotating line was strobed and asked observers to report how many lines they saw at any one time. They derived the duration of visible persistence from the slope of the linear functions relating the strobe rate and the number of lines observers reported. Estimates of the duration of visible persistence ranged between 133 and 144 msec.

The most significant difficulty with these procedures for estimating visible persistence is that the observer must count the number of perceived lines. To determine the visible persistence of stroboscopic stimuli that approximate real motion, we must estimate the persistence of closely spaced stimuli. This requires counting a large number of closely spaced lines, where both the spacing and the number make counting impractical. Alternatively, the classical procedure (Newton, 1720; Allen, 1926) for estimating persistence of an object in real motion (revived by Burr, 1980) utilizes the length of the object's blur streak to estimate visual persistence. While it avoids the counting problem, this method still requires the subject to estimate the size of a rapidly moving object.

A second problem occurs when the spatial position of the stimuli in stroboscopic or real motion is uncertain. In this paradigm (Efron and Lee, 1971) the experimenter has no control over where or when the count of visible lines occurs. Further, the experimenter does not known during what fraction of the trajectory the reported number of lines is visible.

Third, the observed duration of persistence and the number of simultaneously visible stimuli are not absolutely constant from trial-to-trial but, like everything else psychologists measure, vary. The stochastic nature of these measures must be reflected in the data collection and analyses procedures. Thus, the observed duration of visible persistence should be represented by a random variable. The explicit treatment of persistence as a random variable in data analysis, and the measurement of its distribution may prove useful for evaluation of potential theories.

We propose here a paradigm and a method of analysis to overcome the problems of counting, of spatial indeterminacy, and of measuring the random variation of persistence. The paradigm is used to extend the range of spatial and temporal conditions over which it has been possible to measure persistence in stroboscopic motion. The analysis is used to obtain estimations of the complete trial-to-trial distributions of persistence in the various conditions.

The paradigm

In our experiments, two vertical lines one above the other, move horizontally in stroboscopic motion in opposite directions over a fixed distance (Fig. 2). Successive positions, are separated by a fixed interval of time Δt and a displacement of Δx to the right for one line and $-\Delta x$ (leftward) for the other. For different Δt and Δx, observers report whether or not all the lines in both paths appear to be simultaneously present. They are instructed to respond "yes" if they perceive a flickering grating composed of all the positions of the lines and to respond "no" if they do not.

To estimate the duration of visible persistence with this paradigm, we assume that each briefly presented stimulus generates a visual response that decays over time. If the first presented stimulus in one row is still visible when the last presented stimulus occurs in the other row immediately above or below it, the observer responds "visible"; otherwise, "not visible". This paradigm determines the proportion of trials on which a stimulus remains visible from the first flash to the beginning of the last flash in a row.

Responses are inherently probabilistic. We assume that they reflect trial-to-trial variability in either or both the temporal waveform of the persistence response and in the subject's criterion for deciding whether the stimulus is visible. The analysis takes into account the probabilistic nature of the data in order to separate the effects of the retinal separation on (1) the mean duration of visible persistence and on (2) the trial-to-trial variation of visible persistence. The analysis does not distinguish between causes of variability, such as fluctuations in the underlying visual response and fluctuations in the threshold criterion.

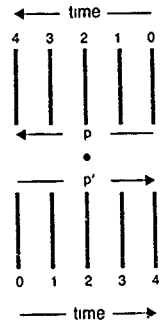


Fig 2 The display for Expts 1–3 two vertical lines were presented in a succession of positions along the paths p and p' as shown above. Each position of the line followed the other by a fixed interval of time, Δt, and was displaced by a fixed distance, Δx, in a constant direction (left or right).

EXPERIMENT 1

Method

Subjects. Data were collected from four observers, including one of the authors (JF). All observers had normal or corrected-to-normal vision.

Stimuli. The stimuli were vertical lines drawn on a HP1310 crt display with a P4 phosphor. The background of the display was illuminated by incandescent lights that produced a background luminance of 0.35 cd/m^2. Subjects viewed the display from a distance of 94 cm and each vertical line subtended 0.235 deg of visual angle (0.386 cm). Each line was displayed for less than 1 msec at the same stimulus intensity. The horizontal and vertical distance between the centers of adjacent raster pixels was 0.0193 cm and each stimulus was composed of a vertical column of 20 raster pixels. Each pixel had a luminance directional energy (cf. Sperling, 1971) of 0.09 cd-sec. This stimulus intensity will hereafter be referred to as the *reference intensity*.

Two vertical lines were presented in a succession of positions, each position following the other by a fixed interval of time, Δt, and displaced to the right (or left) by a distance, Δx, as shown in Fig. 2. One of the vertical lines was presented with its bottom 0.12 deg above a fixation point and extending upward for 0.24 deg. The other vertical line was presented symmetrically 0.12 deg below the fixation point. The two vertical lines were presented in the same horizontal positions, differing only in a spatial shift in the vertical direction and in the temporal order of presentation. On each trial, the direction of motion of the upper line was randomly chosen, the lower line moved in the opposite direction.

Subjects were instructed to stare at the fixation point for the duration of each stimulus presentation. The fact that the two vertical lines moved in opposite directions helped subjects to keep their gaze on the fixation point and discouraged them from tracking the stimulus with their eyes. Making any eye movement during the display would often cause it to appear distorted (see Farrell, Putnam & Shepard, 1984) and subjects quickly learned to suppress eye movements.

Across trials, stimuli differed in the distance between successive lines, Δx, the time interval separating the successive lines, Δt and the total number of lines that were presented, N. The distance, Δx, separating successive positions of each vertical line was either 0.12, 0.18 or 0.36 deg of visual angle. The length of the horizontal path of each vertical line was equal to the product of $(N-1)$ and Δx. For example, when Δx was 0.12 deg of visual angle, N, was 13, 16 or 19 in order to obtain path lengths corresponding to 1.44, 1.80 and 2.16 deg, respectively. Similarly, when Δx was 0.18 deg, N was 9,

11 or 13 for the three respective path lengths. And when Δx was 0.36 deg, N was either 5, 6 or 7 for path lengths equal to 1.44, 1.80 and 2.16 deg, respectively. The stimuli were presented in separate blocks of 120 trials for each condition of path length.

Procedure. The subject initiated a trial by pressing a response key. After 600 msec, two vertical lines were presented in a succession of positions, one line beginning from the left of the fixation point and proceeding to the right and the other line beginning from the right of the fixation point and proceeding to the left. At the end of each trial, the subject pressed one of two response keys to indicate whether or not all successive presentations of the lines on both trajectories appeared to be simultaneously visible. Subjects were specifically instructed to respond "yes" if they perceived a flickering grating composed of all the positions of the lines above and below the fixation point and to respond "no" otherwise.

An experimental session consisted of three blocks of 120 trials corresponding to the three different path length conditions. Within each block of trials, each condition of spatial separation Δx was presented 40 times. The 40 repetitions were presented within two interleaved staircases. The total 120 trials resulting from the product of the 3 Δx, the 20 repetitions per Δx, and the 2 staircase conditions were presented in a random order.

The interstimulus interval was controlled by a modified up-down staircase (Levitt, 1970). The starting value of the interstimulus interval (ISI) in the first experimental session was 50 msec. If the subject responded "no", the Δt was decreased by 2 milliseconds and this new Δt was stored for the next presentation of this staircase. If the subject responded "yes" for *two* presentations of the same stimuli, the ISI was increased by 2 msec and this new ISI was stored to be presented later in the pre-arranged random sequence of trials. The staircase procedure adjusts the temporal separation so that 71% of the time the N successively presented stimuli appear to be simultaneously present. This same procedure was repeated for another interleaved staircase. The complete set of data provided by the two interleaved staircases allows us to estimate psychometric functions for each condition of spatial separation.

In subsequent experimental sessions, the initial value of the Δt was set equal to the estimated 71% threshold from the earlier sessions plus or minus a random number between 1 and 10 msec. The Δt in subsequent sessions was increased or decreased by a number that was proportional to the slope of the estimated psychometric function to insure that the psychometric function was sampled by at least 4 equally-spaced intervals.

All subjects participated in a minimum of 3 experimental sessions: one subject completed 6 sessions, two subjects completed 4 sessions and one subject completed 3 sessions.

Results and discussion

Method of analysis. Our analysis rests on the assumption that a briefly presented luminous line generates a visual response that decays over time, and that after some time the visual response reaches a threshold below which it is no longer visible. We make no assumption about the shape of the visual response; we simply assume that as long as the visual response generated by the stimulus is above threshold, the stimulus will appear to be present. If subjects report that all N lines appear to be simultaneously present, then we assume that, for some instant during that particular trial, the visual responses generated by the N lines were all above threshold. As a practical matter, from the subject's point of view, the question of N visible lines reduces to the simultaneous visibility of the first and last line. No subject reported that the first and last lines of a trajectory were visible but some interior line had vanished.

Let the observable time interval during which the image of all N lines are visible (i.e. above threshold) be a random variable, D. As noted earlier the random variability in D may be the result of threshold variability in the decision stage, variability of the decay function, or other random effects (noise). At the outset, we assume the distribution of D to be normal with mean τ and variance σ^2. This assumption is directly tested in the process of data analysis. For given values of Δt and N, we wish to find $p(\Delta t, N)$, the estimate of the probability that the first and last lines will appear to be visible simultaneously. $p(\Delta t, N)$ is equal to the probability that the first line has not decayed below threshold during the time interval $(N-1)\Delta t$ separating the onset of the first and last stimulus, i.e.

$$p(\Delta t, N) = Prob[D > (N-1)\Delta t]$$
$$= 1 - \Phi\{[(N-1)\Delta t - \tau]/\sigma\}. \quad (1)$$

where τ is the mean duration of visible persistence over trials, σ is the standard deviation of the duration of visible persistence over trials, and $\Phi[\tau, \sigma]$ is a cumulative normal distribution with mean τ and variance σ^2.

Maximum likelihood estimates of τ and σ were computed for each subject and each Δx. The estimations were performed using the numerical procedure STEPIT (Chandler, 1965) to maximize the likelihood that the data were generated by equation (1). To see how well the estimated mean τ and variance σ of persistence duration represent the data, the estimates were used to predict the frequencies of "visible" responses for each subject in each condition of Δx for each individual Δt reached by the staircase. Each of the 12 estimated normal distributions effectively predict the response probabilities. We cannot reject the predictions on the basis of a basis of a χ^2 test at $p < 0.05$ for any subject in any stimulus condition.

Data for each condition and each subject are shown in Fig. 3. The estimated mean duration of visible persistence, τ, and the estimated standard deviation, σ, of persistence duration are plotted as a function of the spatial separation, Δx, for different values of N. There are several interesting aspects of the data. First, Fig. 3 shows that the mean duration of visible persistence increases with the distance separating the successive stimuli, Δx, for all four subjects. This result re-affirms the basic finding reported by (Farrell, 1984).

Second, Fig. 3 shows that the mean τ and standard deviation σ of the visible persistence duration generated by a briefly presented stimulus do not vary with the number of stimuli, N, that are successively presented. The mean and standard deviation depend only on the distance separating successive stimuli, Δx. This result is also consistent with previous findings. Efron and Lee (1972) and Farrell (1984) observed that the number of successive stimuli that appear

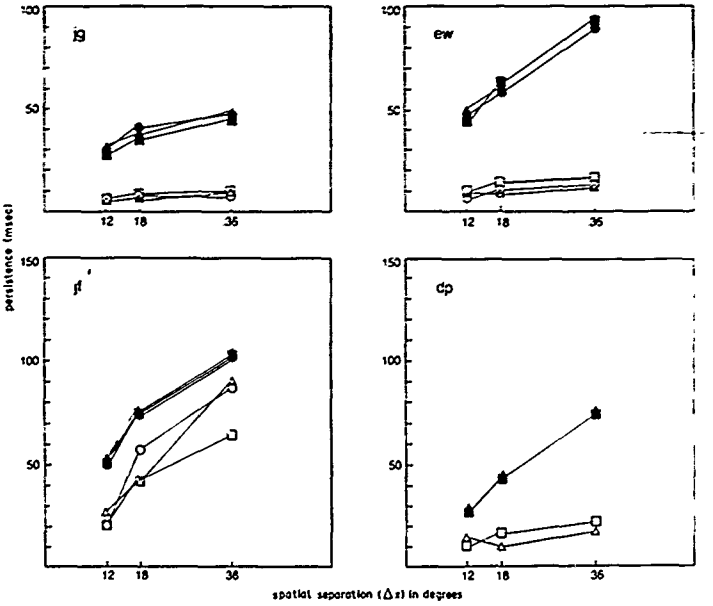


Fig 3 The estimated mean (solid symbols) and standard deviation (open symbols) of the visible persistence of a briefly presented visual stimulus plotted as a function of the distance separating the stimulus line from other stimuli that occur later in time with the length of the stimulus path as the parameter Circles, triangles and squares correspond to stimulus paths of 1 44, 1 80 and 2 16 deg visual angle, respectively Each panel represents data from one subject

to be simultaneously visible trades off with the temporal interval that separates successive stimuli.

The retinal eccentricity of each successively presented stimulus is proportional to $(N-1)\Delta x$. Therefore, the invariance of persistence with N and, consequently, with eccentricity indicates that the duration of visible persistence does not vary with the eccentricities that were investigated (0.7, 0.9 and 1.1 deg). This result suggests that, over the local retinal region investigated, the duration of visible persistence is constant for a given spatial separation Δx. The result does not imply, however, that the retinal eccentricity of a stimulus might not influence the duration of visible persistence if it were varied over a wider range (cf. DiLollo & Hogben, 1985).

Finally, Fig. 3 shows that the variability of persistence duration increases with retinal separation for one of the four subjects (JF). As noted earlier, the individual differences in the variability of the duration of visible persistence across trials may reflect changes in the subjective threshold criterion or changes in the underlying visual response.

EXPERIMENT 2

In the previous experiment we found that for all subjects the mean duration of visible persistence increased with the distance separating the successive stimuli and, for one subject, the variability of persistence duration also increased with the spatial separation. This result is consistent with previous studies that used different experimental paradigms for estimating the duration of visible persistence (Allport, 1968, 1970; Efron & Lee, 1971). These previous studies have not reported limits to the increase of persistence duration with spatial separation. Nonetheless, it seems reasonable to assume that there is both a minimum and maximum duration of visible persistence. In order to place bounds on the increase in persistence duration with spatial separation, we conducted a second experiment and estimate the duration of visible persistence over a wider range of spatial separations.

Method

Subjects The same four observers who participated in the first experiment (EW, DP, JG and JF) served as subjects in this experiment.

Stimuli. As in the previous experiment, the stimuli differed in the distance between successive lines Δx the time interval separating the successive lines Δt and the total number of lines that were presented, N. The number of lines (N) was 25, 13, 9, 7, 5, 4 or 3 for Δx corresponding to 0.06, 0.12, 0.18, 0.24, 0.36, 0.48 or 0.72 deg visual angle, respectively. The lines were displaced over a total path length of 1.44 deg. All other aspects of the stimuli were identical to Expt 1.

Procedure. Each experimental session consisted of two or three blocks of 280 trials. Within each block of trials, each Δx was presented 40 times. The 40 repetitions were separated into two staircase conditions. The 280 trials were arranged in a random order of presentation.

One observer viewed 6 blocks of trials in two separate experimental sessions, another observer viewed 4 blocks of trials in two separate sessions, and two observers viewed 3 blocks of trials in a single experimental session. Observers rested between blocks of trials.

As in the previous experiment, two interleaved random staircases were used to distribute the data around a 71% threshold criteria. Depending on the subjects response, the temporal separation was adjusted such that 71% of the time the N successively presented stimuli appeared to be simultaneously present. The complete data set can then be used to estimate psychometric functions for each condition of spatial separation.

Results and discussion

As in the previous analysis, we assume that the probability that observers will report that the N successive lines appear to be simultaneously present is given by equation (1). Again, using the maximum likelihood procedure, we estimated the values of τ and σ that maximized the match between the predicted and the observed response probabilities for each observer, Δx, and for all values of Δt reached by the staircases.

Figure 4 shows the estimated mean τ and standard deviation σ of persistence duration plotted as a function of the distance Δx for each of the four subjects. Of the 28 estimated normal distributions, only one would be rejected by χ^2 at $P < 0.05$. As in the previous experiment, we found that over a limited range of spatial separations the mean duration of visible persistence increases with spatial separation. In addition, we found that for three of the four subjects (EW, JF, JG), the mean duration of visible

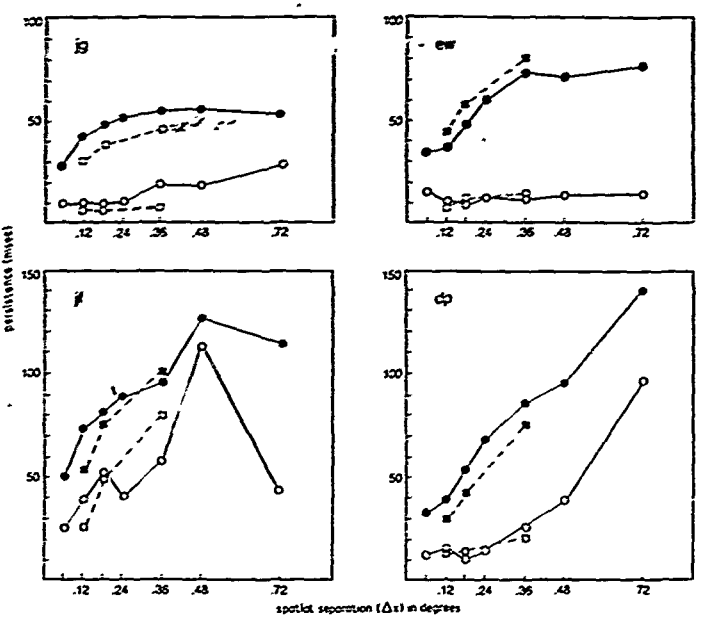


Fig. 4. Estimates of persistence duration plotted as a function of the distance separating successive stimuli for each of the four subjects. Solid circles represent estimates of the mean persistence duration and open circles represent the standard deviation of mean persistence durations. Corresponding estimates of visible persistence derived from Expt 1 are plotted as squares.

persistence approaches a maximum (asymptote) value at the larger spatial separations. The fact that the duration of visible persistence approaches a maximum value at large spatial separations suggests that the mechanism by which the visual system modulates the duration of visible persistence operates over small spatial separations.

Figure 4 also shows that the variability in the duration of visible persistence increases with spatial separation for three of the four subjects (DP, JF, JG) and that the variability is greater at large spatial separations. Most theories of persistence would predict a correlation of τ and σ. For example, if the slope of the decaying visible persistence were to decrease over time, any variability in the threshold criteria for visibility would have greater effects at longer persistence durations. The variability in the persistence estimates for large separations is substantial, however, particularly for subjects JF and DP. This result reduces our confidence in the persistence estimates for large spatial separations.

Finally, Fig. 4 shows the mean and standard deviation of persistence duration estimated from the results obtained in Expt 1. The estimates obtained from Expt 1 are based on stimulus conditions in which the number of successive stimuli, N, varied. The estimates obtained from expt 2 are based on stimulus conditions with constant N. Despite these differences, the mean persistence durations measured in the two experiments fall within the variability in persistence duration for each condition of spatial separation.

EXPERIMENT 3

In the previous experiments, we were able to estimate the mean τ and the variability σ of the duration of visible persistence of a briefly presented visual stimulus as a function of the distance, Δx, separating that stimulus from

other stimuli that occur later in time. We found that the estimated persistence duration τ increases with Δx, and, for 3 of 4 subjects, so does σ. We interpret the average duration τ as the time during which the response to a stimulus remains above a fixed threshold. In the following sections of this paper we examine the implications of the empirical findings in terms of more formal models. To simplify our analysis, we consider only expected values and, for the time being, we ignore variability.

The results discussed thus far may be interpreted in terms of two types of models. In one type of model the shape of the actual temporal response depends on nearby stimuli. For example, the presence of an adjacent stimulus may increase the rate of decay of the response (see Fig. 5a). In a simple exponential system this can be interpreted as a reduction in time constant. We will call this type of model the *rate of decay* model. In the second type of model, the shape of the temporal response may be invariant, only its amplitude is reduced by the presence of adjacent stimuli (see Fig. 5b). We will refer to this type of model as the *gain* model. The rate

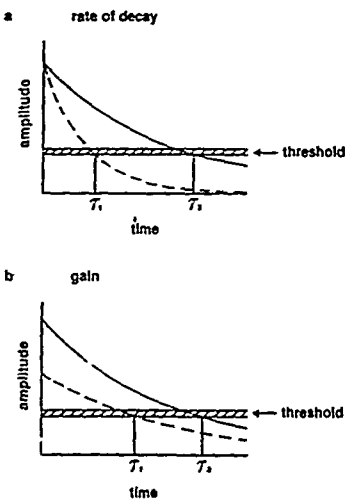


Fig 5 Hypothetical mechanisms for modulating the duration of visible persistence (a) The rate of decay model assumes that the presence of an adjacent stimulus increases the rate of decay, and therefore the shape, of the temporal response (b) The gain model assumes that the shape of the temporal response is invariant, only its amplitude is reduced by the presence of adjacent stimuli

of decay model places no constraint on the shape of the temporal response which can vary with the presence of adjacent stimuli. The gain model constrains the shape of the temporal response to be invariant and, therefore, separable from the influence of adjacent stimuli. In the sections that follow we explore the extent to which a gain model can account for the influence of adjacent stimuli on the duration of visible persistence. We first consider a more formal model of subjects' performance and then describe an experiment to address this issue empirically.

Let us denote the visibility at time t due to a stimulus with intensity I presented at time $t = 0$, $v_{\Delta x}(I, t)$. As before, Δx represents the spatial separation of adjacent stimuli. For simplicity we assume that v is monotonically decreasing (decaying) in time and monotonically increasing with luminance. The value of visibility, v, is used by the subjects to make a decision about the presence of a visible stimulus at each location.

An implicit assumption underlying our data analysis thus far is that the stimulus is visible whenever v was large enough to exceed a fixed threshold c. The estimation of the visible persistence from the results of Expts 1 and 2 amounted to estimating τ_c such that

$$v_{\Delta x}(I, \tau_c) = c. \qquad (2)$$

The estimate of mean persistence duration τ_c, or simply τ, as a function of Δx and Δt for a constant value of luminance I was justified to the extent the criterion c is independent of Δx and Δt, i.e. that the stimulus is visible whenever the visibility function v is greater than a fixed threshold value, c, and that c is constant for all Δx and Δt.

The gain type of model is based on the idea that the distance separating successive stimuli affects only the *amplitude* of the underlying temporal response, v. The amplitude of the response is likely to depend on the stimulus luminance as well. Therefore, in order to develop a gain type of model, it is necessary to separate the effects of luminance and the effects of spatial separation on the temporal response, v. To do this, we first examine the effects of luminance on estimates of persistence duration.

Method

An experiment to test the effects of luminance was performed. The method, apparatus, pro-

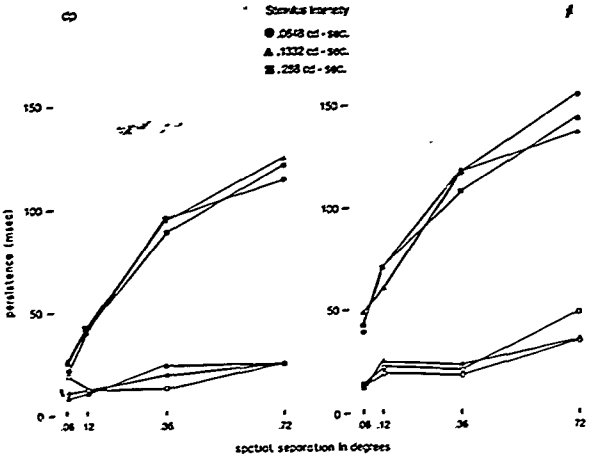


Fig. 6. Estimated mean (solid symbols) and standard deviation (open symbols) of persistence duration plotted as a function of the distance separating successive stimuli with stimulus intensity as a parameter. Stimulus intensity is specified as a fraction of the reference intensity (see *Stimuli* for Expt 1).

cedure and paradigm were identical to that of Expts 1 and 2 except that, in any given trial, the luminance of the briefly presented line was 0.72, 1.48 or 3.2 times the reference intensity (see *Stimuli* in Expt 1) and the spatial separation of successively presented lines was 0.06, 0.12, 0.36 and 0.72 deg visual angle. The distance between the centers of adjacent pixels was 0 0193 cm in both the vertical and horizontal direction.

Results

In Fig. 6, the estimated means and standard deviations in visible persistence are plotted as a function of spatial separation with stimulus luminance as a parameter for the two subjects, JF and DP. Figure 6 shows that there were no systematic effects due to stimulus luminance. Differences in the mean duration of visible persistence due to stimulus luminance are small and inconsistent and can be explained by the variability of persistence duration: for each condition of spatial separation, the mean duration of visible persistence estimated for a stimulus of a given luminance value falls within the standard deviation of the persistence durations estimated for stimuli presented in any of the three luminance values. The results of this experiment can be described very simply: the persistence estimates are invariant with respect to 1:4 luminance changes

Discussion

The visibility criterion depends on peak visibility. The goal of the following discussion is to examine how well the data can be accounted for by a model that assumes that the visibility of a briefly presented line can be represented as a product of three different functions depending on luminance, distance and time, respectively. We begin by noting that brighter flashes do not persist longer than dim flashes. This result suggests that the criterion c depends on luminance in the same manner as does the visibility v. In other words, the results are consistent with the hypothesis that criterion is a threshold defined in terms of a fixed fraction of the initial amplitude of the visual response at time $t = 0$ which is, in turn, a monotonically increasing function of the maximum luminance.

We can express the notion of a relative criterion that is determined by the brightest stimulus on a given trial formerly as follows Let l_m be the luminance of the brightest, briefly presented stimulus line on a given trial Another stimulus line presented with luminance l on the same trial will be visible after a delay t if:

$$v_{\Delta x}(l, t) \geq c(l_m), \qquad (3)$$

where c is a monotonically increasing function of the maximum luminance.

Separability of luminance and distance effects. The threshold criterion c is, as before, assumed to be independent of the spatial and temporal stimulus parameters, Δx, Δt. At the visibility threshold, the inequality (3) becomes an equality and we can divide both sides of this equation by the threshold c. The resulting ratio $v/c = 1$ is independent stimulus luminance. Consider trials where all stimuli are presented with the same luminance I. Then $I = I_m$, the ratio v/c can be used to define a new function w:

$$w_{\Delta x}(t) = \frac{v_{\Delta x}(I, t)}{c(I)}; \tag{4}$$

which does not depend on the luminance level. We have already defined w to be independent of luminance at threshold. If we further assume that w is independent of luminance above the threshold, then the visibility v can be written as a product of two functions:

$$v_{\Delta x}(I, t) = c(I)w_{\Delta x}(t), \tag{5}$$

where c is a monotonically increasing function of luminance, I, and w is a monotonically decreasing function of t and increasing in Δx. Thus v is a separable function of luminance and another function w that depends on time and separation. Note that the function w is independent of luminance and embodies the dependence of persistence on spatial separation Δx.

Separability of time and distance in a gain control model. With this framework at hand, we are ready to formalize the assumption underlying the gain type of model. In that model, the presence of adjacent stimuli only modulates the magnitude of the response. That is, the function w itself can be separated into a product of two functions, gain g, and temporal response h, as follows:

$$w_{\Delta x}(t) = g(\Delta x)h(t);$$

and the visibility function can be written as

$$v_{\Delta x}(I, t) = c(I)g(\Delta x)h(t). \tag{6}$$

The separability of time, distance and luminance expressed in equation (6) predicts that a decrement in luminance, could completely compensate for a corresponding increment in separation Δx. Alternatively, a decrease in the visibility due to small spatial separation can be compensated by an increase in visibility with luminance. Experiment 4 was aimed at discovering the relationship between spatial separation and luminance. If we know how the amplitude of the visual response changes with luminance, and we know how luminance and spatial separation trade-off in determining the duration of visible persistence, then we can derive how the amplitude of the visual response changes with spatial separation.

EXPERIMENT 4

Experiment 4 tests the extent to which the gain type of model holds and thereby yields more information on the temporal response, h. The approach is based on the measurement of a trade-off between the function of luminance, $c(I)$, and the function of separation, $g(\Delta x)$. Since neither $c(I)$ or $g(\Delta x)$ depend on Δt [i.e. they are separable from $h(t)$], we investigated the effects of luminance and spatial separation when $\Delta t = 0$.

Method

Subjects. The same four observers who participated in the previous experiments (EW, DP, JG and JF) served as subjects in this experiment as well.

Stimuli. As in Expt 2, the stimuli consisted of two sets of vertical lines presented 0.12 deg above and below a fixation point (see Fig. 2). In fact, the stimuli were equivalent to the stimuli in Expt 2 with the following exceptions. Rather than present the lines successively, the lines were presented simultaneously. In addition, the intensity of each line was varied as a function of the position of the line: across a row of vertical lines, the intensity of each line decreased exponentially with stimulus position as illustrated in Fig. 7. Let I_n be the intensity of a line in position n

Fig 7 The display for Expt 4 two sets of vertical lines were simultaneously presented above and below a fixation point The height of each vertical line represents stimulus luminance which decreased exponentially with stimulus position

The intensity of the line in the leftmost (or rightmost) position, I_1, was initialized to the reference intensity (see *Stimuli* in Expt 1). The intensity of the line in a position to the right (or left) of n, I_{n+1}, was I_n^{α} where α is the slope of the exponential decrease. On each trial, the direction of the exponential decrease in intensity (left-to-right or right-to-left) of lines presented above the fixation point was chosen randomly; the intensity of the lines below the fixation point decreased exponentially in the opposite direction. The spatial separation δx is varied by increasing n over a range of 3–25 as in Expt 2.

Procedure. The subject initiated a trial by pressing a response key. After 600 msec, the stimuli were flashed for 1 msec. At the end of each trial, the subject pressed one of two response keys to indicate whether or not all the vertical lines above and below the fixation point were visible. Subjects were instructed to use the same criterion for visibility that they used in the previous experiments. subjects were to respond "yes" if they perceived a grating composed of all the lines above and below the fixation point and to respond "no" otherwise.

At the beginning of each session, subjects repeated 280 stimulus trials from the previous experiment. These 280 trials served to remind subjects of the visibility criterion used in previous experiments and to encourage them to use the same visibility criterion in this experiment. Subjects then viewed 3 blocks of trials, each block consisting of 160 trials Subjects rested between blocks of trials.

Across the three blocks of trials, each condition of spatial separation Δx was presented 60 times. The 60 repetitions were presented within two interleaved staircases. The total 480 trials resulting from the product of the 7 Δx, the 60 repetitions per Δx, and the 2 staircase conditions were presented in random order.

The rate of the exponential decrease in stimulus intensity α was controlled by a modified up-down staircase The starting value of α was 0 99 If the subject responded "yes", α was decreased by 0.01 and this new α was stored for the next presentation of this staircase. If the subject responded "no" for two repetitions of the same stimuli, α was increased by 0 01 and this new α was stored Under the assumption that α is a normally-distributed variable, the staircase procedure converges to the α for which 71% of the time all the n lines are visible to the observer for each condition of spatial separation, Δx. All the data were used to estimate the entire psychometric functions.

Results and discussion

Psychometric functions relating the probability of reporting that all n lines were simultaneously visible to the relative intensity of the dimmest line were calculated for each subject and each condition of spatial separation, Δx. In Fig. 8, the relative intensity of the dimmest line (expressed as the normalized ratio of the minimum and maximum line intensities) accompanying 50% response probabilities is plotted as a function of the spatial separation for each subject. As Fig. 8 shows, the relative line intensities required for all n lines to appear to be visible decreased with spatial separations up to 0.24 deg of visual angle. For larger spatial separations, the relative line intensities required to see all n lines do not vary systematically and therefore we conclude that the intensities are independent of spatial separation

The results of Expt 4 can be interpreted in terms of the gain control model. In particular, considering the form of the visibility function v given by equation (6) we set $t = 0$ and interpret Expt 4 as finding values of the dimmest, N-th line l_N for each Δx such that

$$c[l_N(\Delta x)]g(\Delta x)h(0) = c(l_1), \qquad (7)$$

where l_1 is the first (brightest) line. There are three unknown functions in this equation c, g and h and our goal is to determine h We do that in two steps. First, we use previous information on intensity scaling to assume a reasonable form for the criterion function c. We then combine the results of Expts 1, 2 and 4 in order to eliminate g.

The criterion function c represents the observers' adjustments to changes in luminance. To proceed with our analysis we need to make an additional assumption about the function $c(l)$ In particular, we assume $c(l)$ to be a power law. This assumption is consistent with at least two empirical considerations. First, the classical scaling data derived from magnitude estimation experiments (Stevens, 1957) suggests that perceived brightness is a linear function of luminance raised to a power. Second, the assumption is consistent with the luminance invariance observed in Expt 2.

Substituting l^{β} for c in equation (7) yields.

$$l_N^{\beta}(\Delta x)g(\Delta x) = c_0 l_1^{\beta},$$

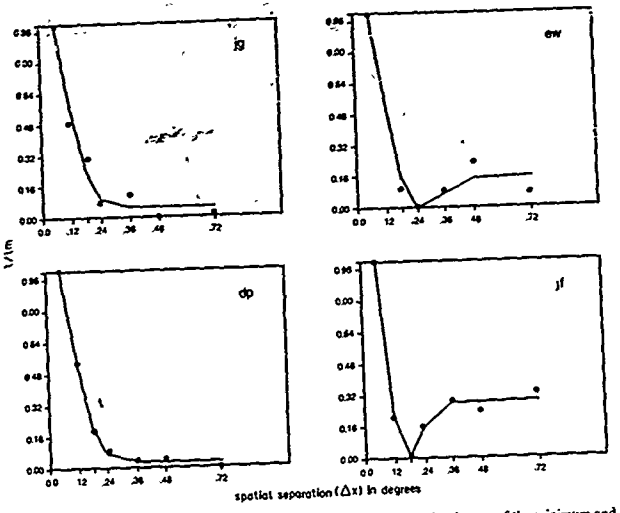


Fig. 8. The relative intensity of the dimmest line (expressed as the normalized ratio of the minimum and maximum line intensities) accompanying 50% response probabilities is plotted as a function spatial separation for each subject. The solid line represents the fit between equation (10) (see text) and the data.

where c_0 is a constant [incorporating $h(0)$]. Taking logarithms of both sides yields the following equation relating Δx and I:

$$\log[g(\Delta x)] = \log(c_0) - \beta \log\left[\frac{I_N(\Delta x)}{I_1}\right]. \quad (8)$$

This equation represents the relationship between two functions of the stimulus separation $I_N(\Delta x)$ and $g(\Delta x)$. Our primary goal is to use the equation (8) to combine the results of Expt 4 with those of the earlier experiments and directly evaluate the shape of the temporal response, h. It is also possible, however, to examine whether there exist plausible gain functions g consistent with both equation (8) and the results of Expt 4. In order to find such a g we first determined a functional form for $I_N(\Delta x)$. While there are many different functions consistent with the empirical constraints on I, we selected the following spatial weighting function generated by taking the difference between two Gaussian functions:

$$\frac{I_N(\Delta x)}{I_1} = A_1\phi(\Delta x, \sigma_1) - A_2\phi(\Delta x, \sigma_2); \quad (9)$$

where $A_i > 0$ are the amplitudes, $\sigma_i > 0$ standard deviations of the positive ($i = 1$) and negative ($i = 2$) Gaussians, respectively, and where ϕ is a Gaussian density function of the form:

$$\phi(x, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma_i}\right)^2};$$

centered at the origin. This type of spatial weighting function seemed plausible because, given the correct parameters, it has been used to describe other spatial interactions including empirically observed receptive fields in monkey and cat retinal ganglion cells (Enroth-Cugel & Robson, 1966). The difference between two Gaussian functions has also been used to approximate psychophysically defined spatial weighting functions (e.g. Schade, 1956; Wilson & Bergen, 1979; Graham, 1980).

The best-fitting parameters to equation (9) were derived for each subject using an iterative fitting procedure (STEPIT, Chandler, 1965) that minimized the squared error between each subject's data and equation (11). The resulting fits, shown in Fig. 8, are quite reasonable. The root-mean-square error for the fits is 0.05, 0.013, 0.058 and 0.025 for subjects JG, DP, EW and JF, respectively.

Since the dependence of the luminance on Δx can be characterized as a difference of two

Gaussian functions then the resulting gain function g, shown in Fig. 8, is also a difference of Gaussians but raised to a positive power β. This function:

$$g(\Delta x) = k[A_1\phi(\Delta x, \sigma_1) - A_2\phi(\Delta x, \sigma_2)]^\beta;$$

where k is a positive constant, appears to be a reasonable reflection of the effect of spatially adjacent stimuli on persistence. According to these results, the width of the *effective field* within which one stimulus line affects the persistence of another is approx. 0.24 deg of visual angle. Since the form of the gain function appears to be reasonable we proceed to use the data from Expt 4 to derive the temporal dependency h. Note that the following derivation is independent of the form of the gain function.

Derivation of temporal dependency. Assuming that the gain model holds, the temporal waveform of the underlying visual response to a briefly presented visual stimulus is embodied in the function h. To evaluate h we need to eliminate g in equation (7). We accomplish that by substituting, in equation (7), the expression for g from equation (8). Empirically, this amounts to combining the results of Expt 4 with those of Expts 1 and 2.

To combine equations (4) and (8) we first take the logarithm of both sides of equation (7), and then solve for g with the result $\log[g(\Delta x)] = \log[b] - \log[h(t)]$. Then, substituting for $\log[g(\Delta x)]$ in equation (8) yields:

$$\log[b] - \log[h(\tau)] = \log(c_0) + \beta \log\left[\frac{I_m}{I}\right];$$

which can be simplified to:

$$\log[h(\tau)] = k + \beta \log\left[\frac{I}{I_m}\right]; \qquad (10)$$

where k is a real constant. To estimate the temporal decay function h consistent with our results can be accomplished by finding a function of τ which is linear in $\log[I/I_m]$.

For each subject and each condition of spatial separation, $\log[I/I_m]$ was estimated by the 50% threshold criteria of psychometric functions relating the probability that the subject responded "yes" (to indicate that all stimuli were visible) to the ratio of the minimum and maximum stimulus luminances, I/I_m. Figure 9 shows $\log[I/I_m]$ plotted as a function of $\log(1/\tau)$ (derived from the data collected in Expt 2) for each subject. The solid lines in Fig. 9 illustrate that the following equation provided a reasonable fit to the data for spatial separations less than or equal to 0.24 deg of visual angle:

$$\log\left(\frac{1}{\tau}\right) = k + a \log\left[\frac{I}{I_m}\right]. \qquad (11)$$

We can therefore conclude that, to the extent that this equation is supported by the data, the gain model cannot be rejected (at least for small spatial separations) and that the decay of visible response has the general form $1/t$. This function might not be a realistic impulse response for a linear system, but it does indicate that the decay of visible persistence is slower than a simple exponential (cf. Rumelhart, 1969; Hawkins & Shulman, 1979; DiLollo, 1984).

Finally, Fig. 9 shows that for larger values of τ (corresponding to stimulus conditions in which Δx was greater than 0.24 deg of visual angle) there seems to be systematic departure from the straight line. This represents the failure of the model to capture spatial interactions over larger separations.

GENERAL DISCUSSION

Persistence is a property of any linear system with limited temporal bandwidth. Usually, the narrower the bandwidth the longer the persistence. Similarly, the more veridical the temporal response of a system is, the less persistence there is. In any sensing system, there is a trade-off between the ability to reproduce the temporal properties of a stimulus (achieved by broad temporal bandwidth and, consequently, short persistence) and the ability to detect the presence of a weak stimulus in the presence of noise (achieved by temporal summation and, consequently, long persistence). There are many situations in which the visual system sacrifices temporal bandwidth in favor of stimulus sensitivity. For example, the time constant of temporal integration is more than two times longer in the dark adapted eye than in the light adapted eye. (Sperling & Sondhi, 1968). We report an instance in which, depending on the spatio-temporal properties of the stimulus, the visual system sacrifices either temporal bandwidth or stimulus sensitivity. When the distance between successive stimuli is small, as in the case of the apparent motion of a single object, the visual system sacrifices stimulus sensitivity in favor of temporal fidelity, preserving the temporal stimulus information and reducing the smear that would otherwise be generated by moving objects (Burr, 1980). When the distance between

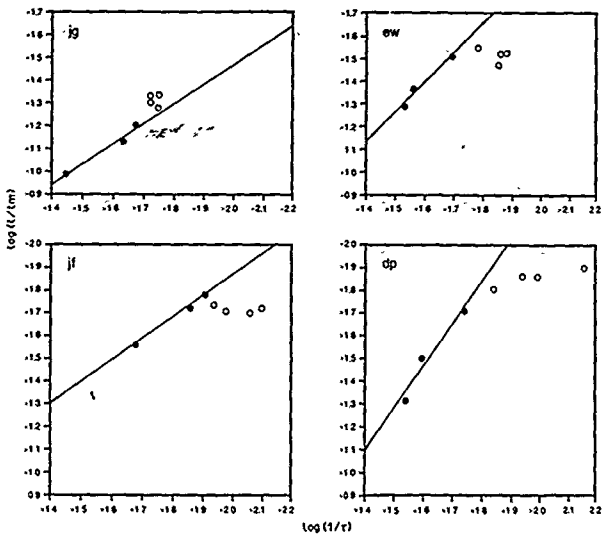


Fig 9. For each subject, the natural logarithm of I_m/I (the ratio of maximum (I_m) and minimum (I) stimulus luminance that accompanied the 50% visibility threshold criteria in Expt 3) is plotted as a function of $1/\tau$ (the reciprocal mean duration of visible persistence estimated from the results of Expt 2) The solid lines represent the linear regression of $\ln[I/I_m]$ on $\ln(1/\tau)$ for the data corresponding to conditions in which adjacent stimuli were separating by distances less than or equal to 0 24 deg visual angle The solid circles falling near the regression line from left to right correspond to spatial separations of 0 6, 0 12, 0.18 and 0 24 deg visual angle, respectively. The unfilled circles correspond to spatial separations greater than 0 24 deg visual angle

successive stimuli is large, as in the case of briefly presented stationary objects, the visual system sacrifices temporal fidelity in favor of stimulus sensitivity, allowing more time to extract the spatial information necessary for object identification.

We consider a simple *gain* model as a possible mechanism for modulating the duration of visible persistence as a function of the distance separating stimuli. In this model, the shape of the underlying visual response is preserved and only its amplitude is modulated by the presence of adjacent stimuli. Our analysis does not assume any particular shape of the temporal impulse response function. We only assume that a briefly presented luminous line generates a visual response that decays over time, and that after some time the visual response reaches a threshold below which it is no longer visible. The effective duration of visible persistence corresponds to the duration that the visual response generated by the stimulus is above threshold. In order to test the gain model, we make the further assumption that the amplitude of the visual response to briefly presented stimuli increases with stimulus luminance and that the effects of spatial separation, luminance and temporal separation on visible persistence are separable. The trade-off we observed between the effects of spatial separation and stimulus luminance on the duration of visible persistence supports the assumption of separability and the gain model The gain model is appealing because it can be realized by mechanisms underlying shunting lateral inhibition (Sperling & Sondhi, 1968).

Acknowledgements—This work was supported by the National Institute of Mental Health, grant 5-T32-MH14267, and by the USAF Life Science Directorate, grant 80-0279

REFERENCES

Allen, F (1926) The persistence of vision *American Journal of Physiological Optics*, 7, 439–457

Allport, D. A. (1968). Phenomenal simultaneity and the perceptual moment hypothesis. *British Journal of Psychology, 59*, 395–406.

Allport, D. A. (1970) Temporal summation and phenomenal simultaneity: Experiments with the radius display. *Quarterly Journal of Experimental Psychology, 22*, 686–701.

Burr, D. (1980). Motion smear. *Nature, London, 284*, 164–165.

Chandler, J. P. (1965). *STEPIT, quantum chemistry program exchange*. Bloomington, Indiana: Department of Chemistry, Indiana University.

Coltheart, M (1980). Iconic memory and visible persistence. *Perception & Psychophysics, 27*, 183–228.

D'Zmura, M. & Lennie, P. (1986). Mechanisms of color constancy. *Journal of the Optical Society of America, A3*, 1662–1672.

Dixon, P. & Di Lollo, V. (1980). Temporal integration in visual memory. *Journal of Experimental Psychology: General, 109*, 75–97.

Di Lollo, V. (1984). On the relationship between stimulus intensity and duration. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 144–155.

Di Lollo, V. & Hogben, J. H. (1985). Suppression of visible persistence. *Journal of Experimental Psychology: Human Perception and Performance, 11*, 304–316.

Dixon, N. F. & Hammond, J. (1972). The attenuation of visual persistence. *British Journal of Psychology, 63*, 243–254.

Efron, R. & Lee, D. N (1971). The visual persistence of a moving stroboscopically illuminated object. *American Journal of Experimental Psychology, 84*, 365–375.

Enroth-Cugell, C. & Robson, J. G. (1966). The contrast sensitivity of retinal ganglion cells of the cat. *Journal of Physiology, London, 197*, 551–566.

Farrell, J. E. (1984). Visible persistence of moving objects. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 502–511.

Farrell, J. E., Putnam, T. & Shepard, R. N. (1984). Pursuit-locked apparent motion. *Bulletin of the Psychonomic Society, 22*, 345–348.

Graham, N. (1980). Spatial-frequency channels in human vision: Detecting edges without edge detectors. In Harris, C. S. (Ed.) *Visual coding and adaptability* (pp 215–262) Hillsdale, New Jersey: Erlbaum

Hawkins, H. L. & Shulman, G. L. (1979). Two definitions of persistence in visual perception. *Perception & Psychophysics, 25*, 348–350

Levitt, H. (1970). Transformed up-down methods in psychoacoustics. *Journal of Acoustical Society of America, 44*, 467–476.

Newton, I. (1720). *Opticks* (reprinted in 1952). New York Dover.

Rumelhart, D. E. (1969). A multicomponent theory of the perception of briefly exposed visual displays. *Journal of Mathematical Psychology, 7*, 191–218

Schade, O. H. (1956). Optical and photoelectric analog of the eye. *Journal of the Optical Society of America, 46*, 721–739.

Sperling, G. (1967) Successive approximations to a model for short-term memory. *Acta Psychologica, 27*, 285–292.

Sperling, G. (1971) The description and luminous calibration of cathode ray oscilloscope visual displays *Behavioral Research Methods and Instrumentation, 3*, 148–151.

Sperling, G. (1976). Movement Perception in computer-driven displays. *Behavior Research Methods & Instrumentation, 8*, 144–151.

Sperling, G. & Sondhi, M. M. (1968) Model for visual luminance discrimination and flicker detection. *Journal of the Optical Society of America, 58*, 1133–1145.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*, 153–181.

Watson, A. B., Ahumada, A. Jr & Farrell, J. E. (1983) The window of visibility: A psychophysical theory of fidelity in time-sampled visual motion displays. NASA Technical Paper 2211, August 1983 (reprinted in the *Journal of the Optical Society of America*, March, 1986).

Wilson, H. R. & Bergen, J. R. (1979). A four mechanism model for threshold spatial vision *Vision Research, 19*, 19–32.

Journal of Experimental Psychology:
Human Perception and Performance
1990, Vol. 16, No. 2, 282-294

Intelligent Temporal Subsampling of American Sign Language Using Event Boundaries

David H. Parish, George Sperling, and Michael S. Landy
New York University

How well can a sequence of frames be represented by a subset of the frames? Video sequences of American Sign Language (ASL) were investigated in two modes: dynamic (ordinary video) and static (frames printed side by side on the display). An activity index was used to choose critical frames at event boundaries, times when the difference between successive frames is at a local minimum. Sign intelligibility was measured for 32 experienced ASL signers who viewed individual signs. For full gray-scale dynamic signs activity-index subsampling yielded sequences that were significantly more intelligible than when every nth frame was chosen. This result was even more pronounced for static images. For binary images, the relative advantage of activity subsampling was smaller. We conclude that event boundaries can be defined computationally and that subsampling from event boundaries is better than choosing at regular intervals.

American Sign Language (ASL) is a gestural form of communication used by the North American deaf and hearing-impaired communities. In free conversation, ASL is as rapid a form of communication as most spoken languages, including English (Bellugi & Fischer, 1972). Over the past decade there have been several investigations of factors related to the transmission of ASL over the existing long-distance communications networks. The problem is to compress a video signal of the signer to the extent that it will fit through a low-bandwidth or low bit-rate communication channel, such as an ordinary telephone line, without greatly disrupting the efficiency of communication. Although previously designed video telephones would suffice for communication, their bandwidth requirements and cost made them impractical. The current public telephone network has transmission limits of 300 to 2800 Hz for analog signals and nominally 9,600 bits per second (bps) for digital signals (ca. 1988).

The authors thank especially the interpreter, scorer, liaison to the deaf community, and advisor on ASL matters, Sue Roberts, for all the help she provided. For their assistance in recruiting subjects, we thank those at the New York Society of the Deaf and Michael Kaufer, program director for the Hearing Impaired at New York University. We also thank all the subjects who participated in the study and who showed genuine interest in assisting the research. We thank August Vanderbeek, who was instrumental very early on in helping to develop the ideas presented here, and Robert Picardi, who overcame all our technical problems.

This work was supported by Grants 85-0364 and 88-0140 from the Air Force Office of Scientific Research, Life Science Directorate, Visual Information Processing Program. In addition, substantial work on the manuscript was completed while David H. Parish was supported by National Institutes of Health Grant EY02934 awarded to Gordon E. Legge at the University of Minnesota.

Correspondence concerning this article should be addressed to George Sperling, Department of Psychology and Center for Neural Sciences, New York University, New York, New York 10003.

Methods of ASL Compression

Spatial Compression

The problem of determining the minimum communication requirements for ASL was posed by Sperling (1978), who subsequently determined that remarkably sparse images could convey intelligible messages (Sperling, 1980, 1981). Sperling, Landy, Cohen, and Pavel (1985) investigated ASL intelligibility versus bandwidth trade-offs by using the most sophisticated spatial image-compression and coding schemes then available. In one condition, subjects were able to interpret signs with normalized intelligibility of .86, in relation to full-bandwidth sequences, even though bandwidth had been reduced to 2880 Hz. Related work by Abramatic, Letellier, and Nadler (1982) with French Sign Language and by Pearson (1981) with British Sign Language also offers the possibility of substantial compression.

The relative success in ASL compression achieved by Sperling et al. (1985) may be attributed to the large amount of redundant spatial information within the ASL signal. Spatial redundancy exists both across individual pixels and across groups of pixels. Individual pixels are redundant when the gray level of one pixel is predictive of the gray level of nearby pixels. Groups of pixels are redundant when cues in one region of the image yield predictions of what should appear in other regions. For example, consider the particular configuration of pixels that yields the form of an arm. To the degree that the hand, elbow, and shoulder are resolvable, other arm pixels are probably unnecessary and are therefore redundant. Accordingly, one may discard some forms of spatial information with the expectation that other, similar information remains. This sort of spatial redundancy provides the basis for the often surprising success of dynamic point-light displays (Cutting, 1978; Johansson, 1973) in conveying rather complex form information. Indeed, Poizner, Bellugi, and Lutes-Driscoll (1981) demonstrated that such displays successfully con-

Atch- 11

very important lexical and inflectional information in ASL and that subjects are quite adept at identifying signs when so presented.

Temporal Compression

In the present study, following Pearson (1981) and Sperling et al. (1985), we investigate temporal redundancy in ASL. By identifying redundant dynamic information, we are able to intelligently subsample an ASL signal in time and thereby retain intelligibility while decreasing the number of transmitted frames. What criteria should guide temporal subsampling? We consider an ASL sequence as a series of events that unfold over time. The subsampled frames should be those that best convey these events. For example, an event might involve moving the left arm from a position parallel to the body to a position perpendicular to the body. It is important to keep clear the distinction between motion and events. *Motions* are continuous and occur at one or more locations simultaneously. An *event* is a connected sequence of frames, the result of segmenting an image sequence. Just as an object in the x, y domain is a grouping together of pixels to which a common label will be applied, an event is a grouping of frames under a common label.

In some instances, the perception of events requires a cognitive component, an interpretation by the viewer (Ebbesen, 1980; Markus & Zajonc, 1985). In other instances, however, there is strong support for the theory that events (or perceptual units) are directly perceived much like motion (Asch, 1952; Heider, 1958; Newtson, Hairfield, Bloomingdale, & Cutino, 1987). Insofar as events can be perceived directly, it implies that there must be consistent stimulus properties that reflect the event structure within a motion sequence.

How does one locate and use such stimulus properties for temporal compression? Interestingly, the problem of dynamic sequence segmentation has been addressed in several domains of study and for a variety of purposes. These include efforts to construct humanlike motion representation schemes (Marr & Vaina, 1980; Rubin & Richards, 1985), the development of motion descriptors for robotics or artificial intelligence (Thibadeau, 1983), and, as noted previously, cause attribution and event perception in the field of social psychology (Heider, 1958; Newtson, 1973). In another ASL study, Green (1984) attempted to locate the boundaries between consecutive signs in a stream of ASL images.

Perceptual Units of Behavior

A technique for determining the location of boundaries of perceptual units that has received considerable attention comes from Newtson (1973) and his collaborators (Newtson & Engquist, 1976; Newtson, Engquist, & Bois, 1977; Newtson & Rindner, 1979; Rindner, 1982). Although Newtson's early work was concerned with determining how the attribution process varies as a function of the unit of perception, he was also concerned with demonstrating that there was, in fact, an objective basis for determining units of behavior. In his procedure, perceptual units were marked by subjects who viewed films and pressed a button connected to a continuous event recorder. Subjects had been instructed to press the button when, "in your judgment, one meaningful action ends and a different one begins" (Newtson, 1973, p. 30). There was a high degree of consistency, both within and between subjects, as to where the boundaries were placed.

In a series of experiments, Newtson and Engquist (1976) showed that subjects were more sensitive to disruptions at the boundaries of behavioral units (breakpoints) than at moments between the boundaries (nonbreakpoints), suggesting that the breakpoints had a high degree of psychological salience. They noted that deletions of frames from ongoing films at breakpoints were more accurately detected than deletions of frames at nonbreakpoints. Moreover, subjects who viewed static presentations of three consecutive breakpoints were more accurate in their descriptions of the action, rated the sequence as more intelligible, and were more accurate at ordering the slides when the slides were shown out of order than subjects who were shown other groups of three frames. In one experiment (Newtson & Engquist, 1976), groups of three breakpoint frames were rated as high on intelligibility as the continuous sequences from which they were drawn. In short, Newtson and his collaborators provided a convincing demonstration that subjectively defined breakpoint frames convey information that is of greater importance to the global percept than nonbreakpoints. (For a review of the role of breakpoints in event perception, see Newtson et al., 1987).

Choosing Significant Frames Automatically

In the present study, we measure the intelligibility of ASL sequences constructed by using a subset of chosen frames from the original sequence. The work of Newtson and his collaborators suggests that in segmenting the sequences, we would do best to retain breakpoints while discarding nonbreakpoints. Although the Newtson procedure for locating breakpoints works well, its usefulness in a real-time communication system is limited, because human observers must select the frames. Given the growing availability of digital image-processing technology, it is natural to digitize the image sequences to be transmitted, to compute which frames represent breakpoints within the sequence, and to transmit the chosen frames and discard the remainder. To implement such a system, we must first consider the physical properties associated with breakpoints: the boundaries of perceptual units.

Physical Characteristics of Breakpoints

Two possible theories could account for the finding that an action stream may be partitioned into discrete units of behavior based on breakpoints (Newtson & Engquist, 1976). Either the particular configuration of components in the scene constitutes a *distinctive state* that identifies the breakpoint, or actions are defined by *state-to-state* changes that are characterized by successive breakpoints. In a test of these two possibilities, Newtson et al. (1977) used a movement notation system, designed for use by choreographers, to code the body

positions of the actors in their sequences. Differences between codings at different points in time describe the position changes of the actor; the surface structure of each sequence is captured. By comparing position changes between successive breakpoints, between breakpoints and nonbreakpoints, and between randomly chosen nonbreakpoints, Newtson et al.'s (1977) results demonstrated strong support for the state-to-state hypothesis but showed no evidence for the distinctive state hypothesis.

The results of these experiments are interpreted to mean that the actors' positions were maximally distinct, within the constraints of the behavior, at successive breakpoints (Newtson et al., 1977, 1987). To achieve maximally distinct body positions at breakpoints, some form of activity must have occurred between breakpoints; that is, the position change cannot happen instantaneously. Newtson et al. (1977, 1987) interpreted their measure of position change over time as an index of movement complexity. This measure, as they pointed out, is related to, but different from, movement magnitude. The important fact that we derive from this work is that some form of dynamic activity must occur between successive breakpoints. This empirical observation supports our subjective impression that some greater-than-average amount of activity occurs between boundaries and that the activity level decreases at boundaries.

Evidence of such activity must be available in the surface characteristics of the sequence. Marr and Vaina (1980) formalized this notion in their state-motion-state representation for segmenting a stream of movement into pieces that can be described independently. They used pauses (described as moments when the parts of a shape are either absolutely or relatively at rest) to segment a motion stream. Pauses occur when the object (or objects) in the scene undergo a change in direction of movement and occasionally occur at other moments in a sequence. This same notion appears in the work of Rubin and Richards (1985), who argued that natural motion boundaries occur at starts, stops, and force discontinuities. Moreover, they provided evidence that human observers have a subjective impression that a significant event has occurred at each of these boundaries. All of these theories imply that one ought to be able to locate event boundaries by tracking some surface characteristic of the dynamic sequence and by searching for frames that correspond to pauses in activity. Rubin and Richards (1985) and Marr and Vaina (1980) theoretically defined ways in which motion sequences might be parsed. In this article we choose a simple realization of these ideas and test its effectiveness for producing intelligible subsampled motion displays.

The Activity Index, $a_A(n)$

The activity index is the fraction of pixels that experience a suprathreshold change in luminance between frames $n-1$ and n. We located event boundaries in ASL sequences by computing this measure of activity between each pair of consecutive frames in each sequence and looking for the local minima. Our activity index was computed by counting the number of pixels that underwent a significant change of gray level between consecutive frames in a sequence; the fraction of such pixels is reported for each frame after the first. This scheme takes advantage of the fact that the percept of motion results from the changing of luminance values over space and time; more activity causes greater numbers of pixels to change from frame to frame. Obviously, it works best when the camera and luminance sources are stable. In a typical ASL sequence, several body parts move simultaneously. Thus, $a_A(n)$ is more indicative of the general level of activity throughout the sequence than of the motion of any particular object.

A sequence consists of N frames; X is the number of rows and Y is the number of columns in each frame. The luminance value at spatial location x, y in frame n, $1 \le n \le N$ is $I(x, y, l, n)$. The number of pixels that change in luminance by more than a threshold amount θ is computed by setting

$$t_A(x, y, n) = \begin{cases} 1 \text{ if } |I(x, y, n) - I(x, y, n-1)| > \theta \\ 0 \text{ otherwise.} \end{cases}$$

The activity index $a_A(n)$ is the fraction of pixels in frame n that experienced a suprathreshold change in luminance:

$$a_A(n) = \frac{\sum_{x=1}^{X} \sum_{y=1}^{Y} t_A(x, y, n)}{XY}.$$

The higher the threshold parameter θ, the smaller the influence of pixels that change as a result of camera or digitizing noise rather than as the result of a moving object.

As an example, we consider a 30-frame sequence in which a white square on a black background moves left and right across the width of a frame sinusoidally. Figure 1 is a graph of $a_A(n)$ for such a sequence. Note that the local minima in Figure 1, where $a_A(n-1) > a_A(n) < a_A(n+1)$, correspond to frames in the original sequence in which the direction of motion is changing (i.e., the peaks and troughs of the sinusoid). Complex movies such as ASL sequences produce complex $a_A(n)$ functions with many more local minima.

Figure 2 shows the $a_A(n)$ function for a 70-frame ASL sequence that shows the sign for the word accident. A drawing

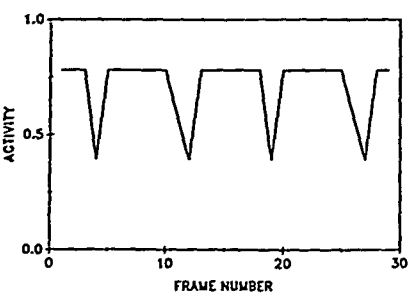


Figure 1 Activity index as a function of frame number for a small square that moves through two cycles of a sine wave across 30 frames. (The local minima correspond to the resting points, or points of change in direction, of the moving object.)

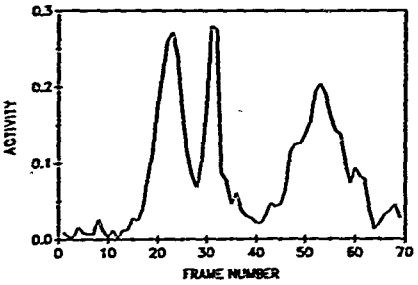


Figure 2. Activity index as a function of frame for a 70-frame ASL sequence that shows the sign for the English word *accident*. (The large amounts of activity at the beginning and end of the function correspond to the action of the signer as she moves into and out of a rest position [arms folded in front]. The two local minima at Frames 28 and 40 correspond to moments when the signer's hands are farthest apart and when they meet in the middle.)

of the sign is given in Figure 3a, and every third frame is shown in Figure 4. In our experiment, the signer assumes a rest position in which her arms are folded in front of her at the beginning and end of the sequence. The same rest position is used for all signs in order to remove any potential ambiguity concerning the beginning and ending of each sign and to ensure that on repeated presentations, the particular frame on which a sign begins or ends does not provide a clue to the identity of the sign. In the sign that produced Figure 2, the signer raises her two hands to either side, closes them into fists, and moves them until they meet in front of her (this is an iconic sign for a collision), and then reassumes the rest position with arms folded.

The $a_1(n)$ function in Figure 2 is instructive for several reasons. First, there is always a high activity value at the beginning and end of each sequence as the signer moves out of and into the rest position. In Figure 2, these peaks occur at Frames 23 and 53. Moreover, at the beginning and end of the sign, the activity index becomes a collection of closely spaced local minima. These frames correspond to the rest position in which luminance noise and slight movements on the part of the signer account for fluctuations in $a_1(n)$, and which might be misinterpreted as significant activities (i.e., these frames might be selected).

Activity-index subsampling. Activity-index subsampling means selecting for presentation only the frames for which $a_1(n)$ has a relative minimum as a function of n. To control the coarseness with which candidate minima are sampled, we introduced a parameter α that specifies the minimum increase in $a_1(n)$ that must occur between consecutively chosen frames, that is, in order to choose both frames f_i and f_j, where $i < j$, the activity index must rise above $a_1(i) + \alpha$ for some frame k, where $i < k < j$. Note that this method of sampling is asymmetric with respect to time; inverting the order of frames may lead to a different selection.

The two large local minima that occur at Frames 28 and 40 in Figure 2 correspond to the point in the sign when the signer's hands are spread apart and to the frame in which they meet. In other words, if we choose frames that correspond to the two local minima at Frames 28 and 40, as well as a beginning and ending frame, we satisfy the criteria for intelligent temporal sampling while reducing the sequence from 70 to 4 frames. These 4 frames are illustrated in Figure 5a.

Rotational and circular motions. Because the activity index relies on changes in activity to indicate event boundaries, rotational or smooth circular motion presents a problem: An activity index may not achieve a significant local minimum during such a motion. Marr and Vaina (1980) recognized the same shortcoming in their state-motion-state representation and proposed to handle these instances by recognizing the occurrence of a confounding movement and dealing with it separately. Our $a_1(n)$ activity index treats rotational and circular motions the same as all others. If we were to discover that subsampled signs never reached some minimum criterion of intelligibility, it might indicate that rotational and circular motion occur often enough within ASL to merit special consideration. However, an informal survey of signs suggests otherwise.

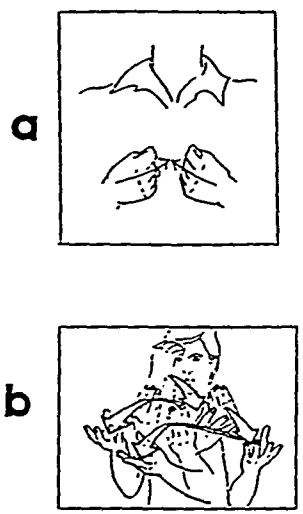


Figure 3. Two illustrated signs showing (a) a simple sign ACCIDENT (from *The Joy of Signing* [p. 102] by L. L. Riekehof 1980, Springfield MO: Gospel Publishing House Copyright 1980 by Gospel Clearing House Reprinted by permission) and (b) a compound sign I-INFORM-YOU-TWO (from *A Basic Course in American Sign Language* [p. 158] by T. Humphries, C. Padden, and T. J. O'Rourke, 1980, Silver Spring, MD: T. J. Publishers Copyright 1980 by T. J. Publishers. Reprinted by permission).

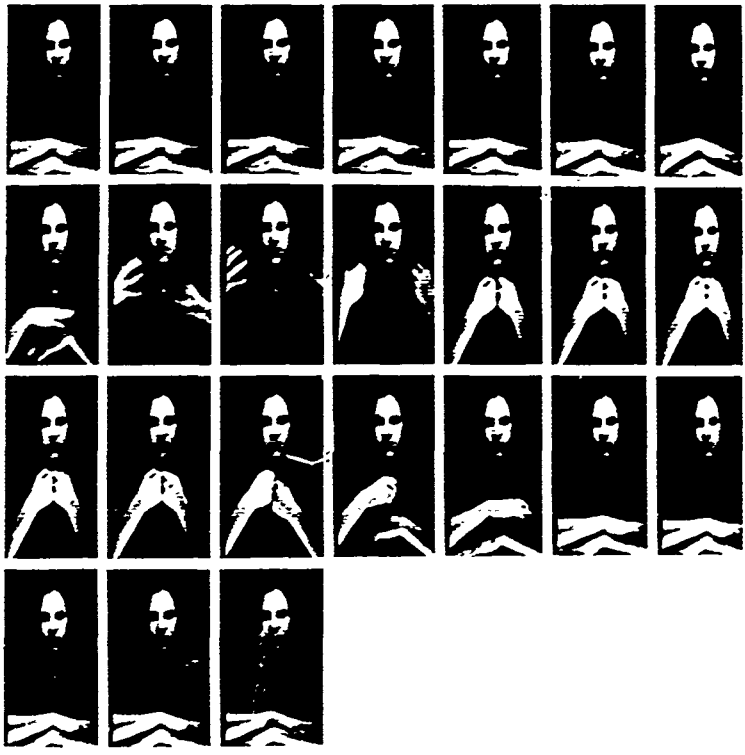


Figure 4. Sequence of digitized images of the sign *accident* (Every third frame of a 70-frame sequence is shown. This is constant subsampling.)

Constant Subsampling

To measure the use of activity-index subsampling, it must be compared with an alternative method of temporal compression. Although the focus of their work was spatial rather than temporal compression, Sperling et al. (1985) and Pearson (1981) used a simple frame repetition, what we here call *constant temporal subsampling*. By this method, every mth frame is chosen from the sequence, where m can take on any value between 2 and the total number of frames in the original sequence. Constant subsampling will be used as the basis of comparison in the present study. Figures 4 and 5b illustrate constant subsampling for m equal to 3 and 23, respectively. In Figure 5b, note that the second frame catches the signer in the middle of a movement.

Dynamic Display Considerations

Having chosen a subset of the frames from a sequence of N frames, how should we choose the duration of each frame to ensure that the displayed sequence retains as much of the rhythmic properties of the original as possible? Temporal constancy is preserved in constant subsampling by choosing the number of repetitions for each frame equal to the constant sampling factor. For example, if every third frame were chosen from the original sequence, each frame in the displayed sequence would be repeated three times

Because the frames chosen from a complex scene via the activity index are not necessarily separated by a constant number of frames, the repetition factor for display must vary according to the location of the chosen frame in the original

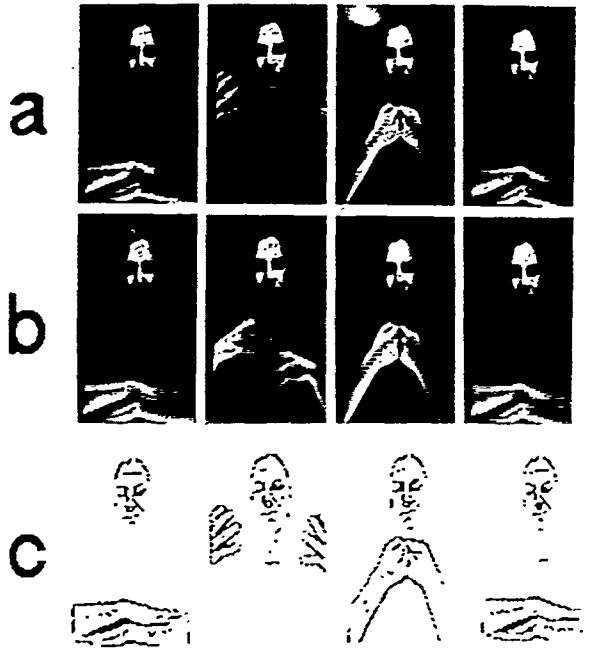


Figure 5 Four-frame representations of the sign *accident*: (a) full gray-scale using activity-index subsampling, (b) full gray scale using constant subsampling, and (c) binary images using activity-index subsampling.

sequence. We repeat each chosen frame (to replace discarded frames) until the next chosen frame occurs. For example, if Frames 1, 5, 15, and 22 were chosen from a 30-frame sequence, Frame 1 is repeated 4 times, Frame 5 is repeated 10 times, Frame 15 is repeated 7 times, and Frame 22 is repeated 9 times, to reach the total of 30 frames that appear in the original sequence. In this method, a different display sequence would be produced from the same sequence of selected frames when played in the forward rather than the time-reversed direction.

Static Presentation

Optimal Number of Frames

ASL-related investigations have, to this point, focused exclusively on the transmission and intelligibility of dynamic images. There are, however, several compelling reasons for studying the intelligibility of ASL when it is presented in static form. Most important, static images are used in most, if not all, ASL textbooks and dictionaries (e g , Humphries, Padden, & O'Rourke, 1980, Riekehof, 1980) Important exceptions are the books produced by Stokoe and his collaborators (Stokoe, 1974, Stokoe, Casterline, & Croneberg, 1976), which use written symbolic notation to convey the motion and hand shape of each sign.

The type of static presentation that is most often seen in standard ASL textbooks is a single-frame image that corresponds roughly to a single English word or expression. Typically, an illustrated signer is presented with overlaid arrows and "strobe" lines to indicate the desired hand, finger, and arm motions. An example of one such illustration, from Riekehof (1980), appears in Figure 3a. For simple signs, especially those that use only one hand, these illustrations are quite efficient. Difficulties can arise, however, for compound signs that require a change in hand shape or for the occasional presentation of complete sentences In these instances, such as for the sign depicted in Figure 3b (taken from Humphries et al , 1980), the many strobe lines and arrows mask the

intended movement and make it difficult for students to replicate the sign.

An alternative to presenting a single frame for each English word or phrase is to present the frames arranged adjacently, as in comic strip format (see Figures 4 and 5). Given the impracticality of displaying the hundreds of frames that may constitute a single sentence, it is necessary to choose a subset of frames for presentation. How does one choose frames in order to convey a sign? Obviously, this is the static analog to the dynamic display that has been addressed earlier and, coincidentally, is a question of great importance to animators and cartoonists. When depicting an action sequence, animators are taught to represent the extremes of the activity first, and then to fill in with in-between frames as needed (Levitan, 1960). In the context of ASL, if frames are chosen to successfully convey the motion when displayed dynamically, do these frames convey the same information when displayed statically?

Spatial-Temporal Compression Trade-Off

Finally, it is useful to investigate interactions between spatial and temporal compression. A practical application would probably combine temporal and spatial compression in order to avoid the degrading effects of removing too much of either spatial or temporal information. Here, we measure intelligibility for both full gray-scale ASL sequences (8 bits/pixel) and for the same sequences made binary with an edge detection scheme. Each image is convolved with a Gaussian-smoothed Laplacian and then thresholded so that 10% of the values are set to black, generally from the dark side of image edges. The result is a binary, line-drawn image (with approximately 0.211 bits/pixel). An example of such a binary sequence is shown in Figure 5c. (These sequences and the nominal data rate were taken from Sperling et al., 1985, Experiment 2, Condition H.)

Method

Subjects

The 32 subjects used in this study were recruited in various places, including the New York Society for the Deaf, the New York University Office for Disabled Students, and word-of-hand among the deaf community. Several fluent hearing ASL interpreters were also used. The mean age of our subjects was 33 years (ages ranged from 18 to 52), and they had been signing for an average of 18 years. Twelve native signers—those who were raised in homes where ASL was the primary language—were included in the study.

Stimuli

The stimulus set consisted of 84 ASL signs, each of which corresponds roughly to a single English word. All signs were taken from Sperling et al. (1985), who filmed, digitized, and applied various image transformations to the signs for use in their study of trade-offs between ASL sign intelligibility and bandwidth. The signer was filmed from approximately 10 ft (approximately 3.05 m) away, so that the upper body and head filled the viewfinder of the camera. During filming, the signer stood behind a screen with a 12 × 18 in. aperture, wore dark clothing, and had dark hair; these conditions ensured that the hand and face of the signer would be highlighted. Each digitized frame was subsequently cropped to 96 × 64 pixels; the signer was centered in each frame so that the area from her waist to the top of her head was visible.

Along with the original full gray-scale (denoted FGS) movies of each sign, we used signs that had been transformed from the FGS to the line-drawn, binary images previously described. Such signs will be referred to as BIN (for binary) signs. Sperling et al. (1985) reported an intelligibility of .911 for BIN signs, normalized against the percentage correct for 96 × 64 pixel FGS signs. At the time of the experiment, four FGS signs were not available under the BIN image transformation, the BIN conditions used four signs that did not appear in FGS conditions, and vice versa. The list of 84 signs used in the study appears in Table 1, divided into the stimulus blocks used in the experiment.

Although the term *frames per second* (fps) is used throughout the remainder of this article to describe the degree of temporal compression, all dynamic stimuli were presented on a system that always displayed 60 fps. In the context of the present study, fps refers to the number of new frames per second, computed by dividing the number of chosen frames by the duration of the original (or the derived) sequence. Frame rate (fps) was varied parametrically. The parameters m and α control the frame rate for fixed-rate and activity-index subsampling, respectively. Four different values were used for each of these two parameters. This manipulation allows us to collect data over a large range of intelligibility. The parameter values and the average number of frames per second for each scheme is displayed in Table 2.

Table 1
Stimulus Blocks

Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8
telegraph	wrong	sit	emphasize	punishment	bear	tobacco	summer
leave	general	cheese	wife	apple	kill	thanks	think
deaf	girl	until	world	uncle	flag	home[a]	talk
finish	short	shoe	our	screwdriver	sorry	flower	wrestling
plan	week	wait	accident	guilty	tree	member	love
ugly	noon	picture	hospital	paper	bread	challenge	read
train	preach	month	friday	understand	behind	pay	start
relax	red	steal	cancel	yesterday	everyday[a]	fun	color
mother	machine	program[a]	because	letter	eye	which	before
jump	improve	spend	boss	bored	cop	grow	alive[a]
		egg[b]			movie[b]	pour[b]	lousy[b]

[a] Signs that appeared only in BIN (binary) conditions. [b] Signs that appeared only in FGS (full-gray scale) conditions.

Table 2
Stimulus Transformations and Frame-Rates

Scheme	α or m	Frames per second
Activity index	0	10.8
Activity index	0.02	8.85
Activity index	0.05	6.75
Activity index	0.1	5.4
Constant	2	30
Constant	4	15
Constant	7	8.6
Constant	1[1]	5.5

Note The α is the parameter that controls the number of samples used by the activity-index sampling scheme; m is the parameter that governs the number of frames chosen by the constant subsampling scheme.

Procedure

The ASL signs were divided into eight groups of 10 signs, each group balanced for difficulty by the criterion of Sperling et al. (1985). The experimental variables included two image types FGS and BIN, two presentation modes, dynamic (D) and static (S), two subsampling schemes, constant and activity index, four frame rates, and 10 stimulus blocks. A full-factorial experiment on these factors would require 320 subjects with only 1 subject in each cell To achieve a more manageable study, we ran four separate groups of subjects, one for each combination of image transformation and presentation mode (FGS-D, FGS-S, BIN-D, and BIN-S).

To make the most efficient use of each subject, the remaining factors within each of the four groups were subjected to a Greco-Latin design in which subsampling scheme and compression factor were fully randomized and order of presentation was partially randomized In other words, each stimulus block of 10 signs was paired one time with every combination of subsampling scheme and compression factor over the course of the experiment For convenience, the combination of subsampling scheme and compression factor is referred to as the stimulus *transformation*, every transformation appeared in each ordinal position of stimulus presentation Order of presentation is only partially randomized, because sequence effects are not balanced in this design. Each subject saw eight complete stimulus blocks, each block having undergone a different transformation (i.e., repeated measures over transformation and stimulus block) A total of 32 subjects were required for a single replication through each of the four 8 × 8 Greco-Latin squares.

Intelligibility test All stimuli were processed with the HIPS image-processing software (Landy, Cohen, & Sperling, 1984a, 1984b) and were presented on a computer-controlled graphics display processor (Adage RDS-3000 image-processing system). Images were viewed on a Conrac 7211C19 monitor, set so that the mean luminance of the display was equal to 55 candela per square meter (cd/m^2) Subjects were seated approximately 1 m from the screen, though they were free to move to their most comfortable distance (Parish & Sperling, 1987, demonstrated that for stimuli whose visibility is impaired by noise, viewing distance, over an extremely wide range, is immaterial)

For all conditions, subjects were required to respond to each ASL presentation with an English gloss for the presented sign Subjects were told that each sequence contained only a single sign and that each sign corresponded, roughly, to one English word In most cases, subjects wrote their responses on an answer sheet In cases in which deaf signers did not possess English skills that were advanced enough to allow them to respond with a written word, they would sign the response to an ASL interpreter who then recorded the English equivalent. The interpreter confirmed these subjects' understanding of the sign by having them either use the word in a sentence or further elaborate on the meaning of the word. Finally, all subjects were told that if they had no idea what the correct answer was, they did not have to respond.

Dynamic presentation. The word *begin* appeared on the monitor, signaling the subject to press any button on a five-button keypad. After the button press, the screen was cleared, and a white cue spot appeared for 0.5 s. This was followed by a 0.5-s blank interval and the presentation of an ASL movie (frame sequence). The sequence was shown once, with the frames repeated as necessary to retain the duration of the original image sequence, and was followed by a blank screen. The word *wait* was displayed until the next sequence was ready for display (2 or 3 s), at which point the word *continue* appeared. While waiting for *continue* to appear, we recorded the subject's response. After the subject's response and after the word *continue* appeared on the screen, the subject was free to press any button to initiate the next trial.

Static presentation. As with the dynamic presentation, the initial button press erased the word *begin* from the screen and caused the stimulus to be presented, though without a cue spot. The frames of each movie were arranged in order by rows and columns, from left to right and from top to bottom Up to seven frames appeared in each row. A sample "page" of 24 frames for the sign *accident* is shown in Figure 4. Shorter pages are shown in Figure 5 for several conditions On presentation, the subject scanned the page and decided on a response Before writing or signing the response, however, a second button press was required to erase the screen After responding, the word *continue* appeared on the screen The next button press initiated the next trial

Results

Scoring

The measure of performance for all subjects and conditions is in percentage correct For some of the signs used in the study, several English responses are considered correct, a result of the historical and regional development of ASL Each subject's answer sheet was scored by a congenitally deaf signer who is fluent in ASL.

Subject Comparison

To assess the general ability of the subjects in this study, we may compare their performance for dynamic sequences with the performance of the subjects from Sperling et al (1985), who viewed similar sequences For the most richly subsampled full gray-scale sequences, which averaged about 20 frames, subjects in the present study averaged 86% correct, nearly identical to the Sperling et al subjects, who averaged about 87% co. .ct. For binary images, subjects in the present study averaged 70%, in comparison with Sperling et al.'s 80% correct Our subjects may have been somewhat less skilled than those of Sperling et al., perhaps a result of the mix of native and nonnative, hearing and nonhearing signers used in the present study, as opposed to the more homogeneous group of deaf signers used by Sperling et al

Main Effects

The data in each of the four Greco-Latin squares, distinguished by the combination of image type and presentation

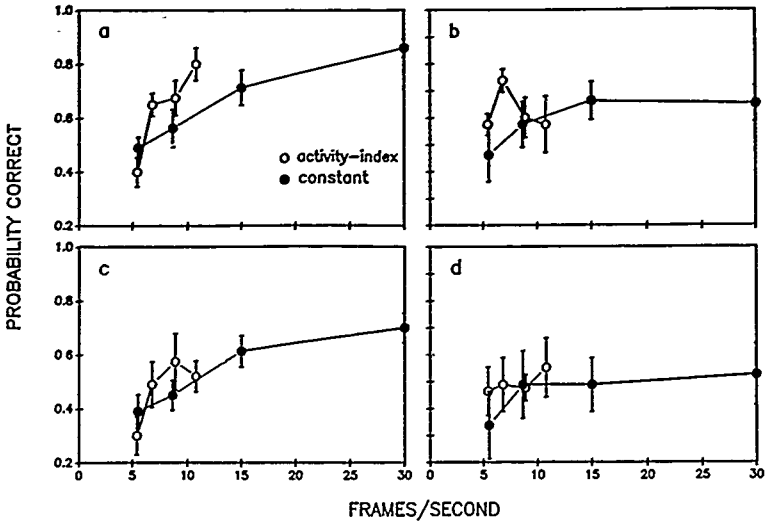


Figure 6. Mean subject performance as a function of the mean number of frames per second for each of the four main conditions of the experiment and for each subsampling scheme. (a) full gray-scale images in dynamic presentation (FGS-D), (b) full gray-scale images in static presentation (FGS-S), (c) binary images in dynamic presentation (BIN-D), and (d) binary images in static presentation (BIN-S) (The open dots on each graph represent performance with activity-index sequences and the solid dots represent constant subsampled sequences. The vertical bars represent the standard error of the mean.)

mode, are displayed in Figure 6. Probability correct is displayed as a function of mean number of frames, averaged across the 8 subjects within each design. These data were subjected to an arcsine transformation in order to decorrelate mean and variance, the arcsine data were used in the subsequent analyses Main effects for each individual Greco-Latin square were evaluated by an analysis of variance.

The four main effects for each Greco-Latin square are subjects, order, stimulus block, and image transformation The most stringent assumption made by the analysis is that the interactions among the four main effects of the Greco-Latin square are negligible (Winer, 1971) Every effort was made to ensure negligible interactions. Subjects were assigned randomly to each cell, and stimulus blocks were balanced for difficulty. There is no a priori reason to assume that there would be significant interactions.

Stimulus transformation, which includes both subsampling scheme and compression factor, was a significant factor for the FGS-D, BIN-D ($p < .01$), and FGS-S ($p < .05$) conditions but not for the BIN-S condition, although there was a trend in the expected direction. Stimulus block was significant for all four designs ($p < .05$). The fact that the stimulus blocks differed from each other indicates that our efforts to equate the blocks for difficulty was not entirely successful. This is not surprising, because Sperling et al. (1985) also found a significant effect of stimulus block despite similar efforts We rely on the fact that throughout the course of the experiment, all stimulus blocks were presented in all conditions, thereby allowing block effects to balance out Finally, there were significant subject differences ($p < .01$) for static presentation of both image types (FGS and BIN).

Subsampling Scheme

The data from the full gray-scale conditions, seen in Figures 6a and 6b, suggest that the activity-index sequences were more intelligible than constant subsampled sequences Ideally, we would have had data from both schemes at the same frame rate to allow us to directly test this hypothesis Unfortunately, the nature of the subsampling schemes prevents such sampling precision. To conduct the test, we used linear interpolation to estimate performance for 6.75 fps for constant subsampling Activity-index data had already been collected at this frame rate. For each presentation format, a t test was computed with the interpolated constant-subsampling data and real activity-index data. Both tests strongly reject the null hypothesis ($p < .01$). For dynamic presentation of binary images (Figure 6c), data interpolated at 8 85 fps for constant subsampling also reject the null hypothesis ($p < .05$).

For dynamic presentation, activity-index performance is estimated, by averaging between points, to be about 8% better than for constant subsampling (for the portions of the curves that overlap). This estimate reflects, in part, the crossover interaction that occurs at the lowest frame rate, in which constant subsampling outperforms activity-index performance. This crossover interaction almost certainly comes from the fact that the activity-index scheme chose frames nearer to the beginning and end of the original sequence when working at extremely coarse sampling rates, whereas constant subsampling chose frames uniformly throughout the sequences. This tendency of activity-index subsampling was due to the large movements that occurred as the signer moved in and out of the rest position, producing the only $a_A(n)$ values that rose above α. If this artifactual performance is discounted—that is, if the beginning and ending rest positions are removed from consideration—the estimate of overall activity-index superiority to constant subsampling increases to 15%.

The form of activity-index performance varies with stimulus presentation. For static presentation of full gray-scale images (Figure 6b), activity-index performance rises above that of constant subsampled images as the total number of frames decreases, the estimated difference in performance rises by nearly 20% when 6.75 fps are displayed. Activity-index performance, however, falls off sharply with fewer and greater numbers of frames per second. Interestingly, for activity-index sequences, FGS-S has a performance maximum at about 7 fps, for constant subsampled sequences, performance with FGS-S improves monotonically with frames per second. In contrast, static presentation of binary images (Figure 6d) produces flat and nearly equal performance for both subsampling schemes, reflected in the nonsignificant transformation factor in the analysis

Discussion

Structure of Events

Although the experiment described here is not a direct test of the validity of Newtson's (1973) definition of breakpoints, there is certainly a close relation between their action-unit boundaries and our basis for dynamic sequence segmentation. Insofar as such a comparison may be made, the results reported here generally confirm findings of Newtson and his collaborators with regard to the perceptual salience of boundaries and their ability to convey critical event information. Indeed, in one condition, ASL sequences that were constructed via the activity index were as intelligible as the original sequence from which the frames were taken, despite a four-fold reduction in the number of frames. This result is similar to an intelligibility-rating result of Newtson and Engquist (1976) and yet, because of the objective nature of ASL intelligibility, is not open to the questions that follow the subjective rating paradigm used by Newtson and Engquist

Direct Perception of Events

A long-standing argument in theories of event perception revolves around the issue of whether events are directly perceived, originating in Asch's (1952) theory that action-defining gestalten appear in the behavior sequence, or whether event perception is more of an interpretive, cognitive process. If action is directly perceived, it must be the case that the cues that give rise to the percept exist in the surface structure of the behavior sequence; that is, a necessary condition for the direct perception of events is that the basis for event structure must appear in the stimulus itself. This is the explicit assumption behind the behavioral segmentation method of Newtson (1973), and it is well supported by the many subsequent experiments by Newtson and his collaborators. If complete event information did not exist in the surface structure of the sequences used in the present experiment, it would have been extremely difficult, if not impossible, to segment our ASL movies into intelligible sequences. At the very least, some higher level, interpretive driver would have been necessary in order to produce compressed images that were more intelligible than those produced by constant subsampling. However, it is clear from the results of our experiment that the necessary event information does reside in the surface structure of sequences.

Even if it is conceded that events are directly perceived and that critical event information is carried by event boundaries, we would expect static presentation of behavior sequences to require a more interpretive process than does dynamic presentation of the same sequences That is, events are usually not directly perceived with static presentation (Newtson & Engquist, 1976). Indeed, for dynamic presentation, the change in surface structure from one moment to the next is immediate, whereas for static presentations, the change must be inferred from an analysis of frames. That these are fundamentally different processes is reflected in the demonstration of left-hemispheric advantage for statically presented signs and the absence of lateral asymmetry for dynamic signs (Poizner, Battison, & Harlan, 1979). Because the inference process would certainly introduce an additional source of error, it seems likely that static images would be less efficient at transmitting the desired information. Accordingly, we note the generally lower intelligibility scores for static images in the present experiment.

The current experiment supports the notion that the basis for event structure appears in the stimulus itself We found that the ASL events isolated by a simple image-based computation seem to agree with subjective impressions of event structure. This does not preclude the possibility that other sources of information, including higher reasoning, can act to modify the interpretation of an event. Nonetheless, our findings bode well for efforts in artificial intelligence directed toward machine interpretation of actions

ASL Primitives

A central component in the traditional study of ASL is the use of ASL primitives. Stokoe (1974) developed a set of primitives to describe signs that are composed of a limited set of movements, hand shapes, and locations of articulation These components are meaningless when taken individually; when combined according to rule-governed constraints, they form the lexical basis of ASL This is entirely analogous to

the function of phonemes in spoken language. A fourth ASL dimension, hand orientation, has subsequently been added to the list of primitives (Battison, 1974).

A somewhat remarkable result is the high intelligibility of the ASL sequences at extremely coarse sampling rates. Even in the most degraded condition, subjects correctly interpreted nearly a third of the signs. An explanation of these findings may stem from the relative importance of the four ASL primitives. Consider that a single frame taken from the middle of a sequence will likely convey information about three of the four primitives—hand onentation, hand shape, and location of articulation. Only motion is lost, or at the very least, severely degraded. The fact that subjects do so well with this limited amount of information reflects the degree to which nonmotion factors play a critical role in ASL intelligibility. Indeed, several studies have shown that the four primitives are not equally perceptible, nor are they equally important for intelligible ASL (Klima & Bellugi, 1979, Tartter & Fischer, 1982).

Image Sequence Compression

Activity index: Dynamic sequences It is apparent from the data and from subjective reports that for low frame-rate conditions, ASL sequences that have been subsampled with the activity index are more intelligible than those subsampled by a constant factor. In the full gray-scale dynamic condition (Figure 6a), activity-index sequences were correctly identified 80% of the time at slightly less than 11 new fps. When constant subsampling was used, the same performance level was not achieved until an estimated 20 to 25 new fps were displayed. At this criterion of performance, the number of to-be-transmitted frames was reduced by a factor of 1.8 to 2.25, roughly a twofold improvement over constant subsampling.

What are the implications of our findings? An 8-bit sequence with frames of 96×64 pixels shown at 30 fps requires 1.47 Mbits/s for full bandwidth transmission, more than 300 times the nominal capacity of the public switched telephone network. The large bulk of compression needed to transmit ASL sequences can certainly come from spatial compression and efficient data-encoding schemes, as demonstrated by Sperling et al. (1985). Nonetheless, sharing the effects of compression among spatial and temporal domains reduces the reliance on spatial compression, thereby reducing the amount of spatial information loss. Furthermore, it is easy to conceive of environments in which it is desirable for dynamic information to be transmitted, or encoded, as efficiently as possible. Intelligent temporal subsampling would have to be included in any such scheme.

The degree to which spatial and temporal compression may be joined depends on the degree of interaction between the two domains. Although activity-index subsampling yields sequences that are more intelligible than constant subsampling for the binary images used in the present study, the overall level of intelligibility, in relation to full gray-scale sequences, was reduced (although this particular comparison is across different groups of subjects). This interaction may suggest that there is limited promise in combining the two forms of compression. Indeed, Sperling et al. (1985) found that extreme spatial compression yielded frames that were temporally decorrelated, so that additional temporal compression was ineffective.

It may be, however, that the particular form of spatial compression used in the present study undermined the success of activity subsampling. Our binary images were constructed by painting 10% of the pixels on the dark side of edges black on a white background, and the selection of pixels to darken might vary with slight changes in the signer's position. The physical representation of the signer within the sequence (i.e., the contours) emerged as a result of the juxtaposition of the black pixels averaged over several frames. Accordingly, a single frame taken from the middle of the sequence may represent the form of a human only very poorly; motion is necessary for the true physical structure of the signer to emerge. By disrupting the temporal characteristics of the sequences, we induced a breakdown of the spatial structure of the signer herself. Naturally, with the loss of spatial structure, intelligibility suffered. A better test of the temporal and spatial interaction would be to use a spatial compression scheme that preserves the structure within a frame without relying on motion cues and spatial averaging that occur between frames.

Rotational and circular motions It was noted in the introduction that signs with rotational or circular motions present a unique problem to the sort of temporal segmentation conducted by the activity index. There are changes in the direction of the moving component (or components) without a corresponding change in velocity or acceleration. Depending on the criteria used to define a rotational or circular motion, between 6% and 15% of the signs used in this experiment could be so classified. By using the stricter criteria for inclusion, activity-index performance was compared with constant subsampling performance within a group of five signs with rotational or circular motions. There was no statistical difference between the two subsampling schemes within this group of signs. As expected, activity-index subsampling presented no advantage over constant subsampling for this group. In addition, intelligibility for the group of five rotational/circular signs was compared with intelligibility for the entire stimulus set. Although the difference was not significant, almost certainly a result of the small number of samples, intelligibility for the rotational/circular signs was, as a group, slightly lower than that of the complete set. Again, this is consistent with our expectations.

Despite the shortcomings of activity-index subsampling that appear when confronted with circular or rotational signs, the effectiveness of this technique is not likely to be greatly affected in any environment that more closely resembles the real world. In a continuous stream of signing, contextual constraints of the conversation will increase the overall intelligibility of the individual signs. Although there is a ceiling effect on many easily interpreted signs, these other, more difficult signs will be made more intelligible. Furthermore, we note that these signs usually represent a fairly small percentage of the total number of available signs.

Application, real-time computation Sperling et al. (1985) demonstrated that telephone transmission of intelligible ASL was feasible, the experiments presented here indicate how this

minimal transmission can be improved by activity subsampling. In order for such a system to be useful to the signing public, the necessary hardware must be relatively affordable, easy to use, widely available, and the processing must be carried out in real time. In the present study, all computations were performed in software and required considerable computing power and time. However, the computations (the accumulation of frame-by-frame differences) were deliberately chosen to be of the kind that are easily embodied in parallel microprocessors. Indeed, we do not see any purely technical obstacles to producing video telecommunications devices that can transmit intelligible ASL over the ordinary switched communications network. Such facilities would have enormous practical significance for the signing deaf and hearing impaired, reducing their isolation from each other and, one hopes, from the hearing community at large.

Static Presentation

Optimal number of frames Two interesting findings emerge from the static presentation conditions. First, it is encouraging to note that even in the most difficult condition, there was still a 30% chance of correctly identifying the presented sign This attests to the robustness of the ASL signal. Second, as noted in the Results section, performance for static presentation of full gray-scale signs declines when there are more than 6 fps in the activity sampling condition. Why?

Subjects reported difficulty with the task of scanning through a page of "printed" ASL frames, although they improved with practice The most common complaint was that there were "too many frames to see what was going on." If it were simply the case that there was an optimum number of frames for each sign, then we would have expected to see evidence of this in both subsampling conditions. Yet, this pattern emerged only for activity-index subsampling. The difference is that although activity-index subsampling chooses critical frames, when the frame repetition factor m is increased in constant subsampling, critical frames are just as likely to be discarded as any other frames. For constant frame-rate sampling, the improved performance that would have resulted at the optimal frame rate is compensated by the loss of critical information It is not just that there is an optimal number of frames, but that there are optimal frames, and that activity index subsampling is one method of discovering optimal subsets of frames.

Automatically generated ASL text. The ability of subjects to "read" static signs and our ability to use digital image technology to produce static text raise an intriguing possibility: messages or even books composed entirely of signed sequences. The automatic production of such static text offers ASL signers an opportunity for veridical representation of ASL conversations that is understandable directly without mechanical aids, such as VCRs. Direct quotes, jokes, announcements, and the like can be communicated with individual expression and intonation. It remains to be determined whether signers could, with practice, become sufficiently proficient at reading signed text to make these possibilities practicalities.

References

Abramatic, J. F., Letellier, P. H., & Nadler, M. (1982). A narrow-band video communication system for the transmission of sign language over ordinary telephone lines. In T. S. Huang (Ed.), *Image sequence processing and dynamic scene analysis* (pp. 314–316). New York: Springer-Verlag.

Asch, S. (1952) *Social psychology.* Englewood Cliffs, NJ. Prentice-Hall.

Battison, R. (1974). Phonological deletion in American Sign Language. *Sign Language Studies, 5,* 1–9.

Bellugi, U., & Fischer, S. (1972). A comparison of sign language and spoken language. *Cognition, 1,* 173–200.

Cutting, J. E. (1978). A program to generate synthetic walkers as dynamic point-light displays. *Behavior Research Methods and Instrumentation, 10,* 91–94.

Ebbesen, E. B (1980). The development of behavior perception. In R. Hastie, T. Ostrom, E. B. Ebbesen, R. S. Wyer, D. L. Hamilton, & D. E. Carlston (Eds.), *Person memory* (pp. 46–69). Hillsdale, NJ: Erlbaum.

Green, K. (1984). Sign boundaries in American Sign Language. *Sign Language Studies, 42,* 65–91.

Heider, F. (1958) *The psychology of interpersonal relations* New York: Wiley.

Humphries, T., Padden, C., & O'Rourke, T. J. (1980) *A basic course in American Sign Language* Silver Springs, MD. T J. Publishers

Johansson, G (1973) Visual perception of biological motion and a model for its analysis *Perception & Psychophysics, 14,* 201–211.

Klima, E., & Bellugi, U (1979). *The signs of language* Cambridge, MA: Harvard University Press

Landy, M. S., Cohen, Y., & Sperling, G (1984a) HIPS. A Unix-based image processing system *Computer Vision, Graphics, and Image Processing, 25,* 331–347.

Landy, M. S., Cohen, Y., & Sperling, G (1984b) HIPS Image processing under Unix, software and applications *Behavior Research Methods, Instrumentation and Computers, 16,* 199–216

Levitan, E L (1960) *Animation art in the commercial film industry* New York: Reinhold

Markus, H., & Zajonc, R (1985) The cognitive perspective in social psychology In G Lindzey & E Aronson (Eds.), *Handbook of social psychology* (3rd ed., Vol 1, pp 137–230) New York. Random House.

Marr, D., & Vaina, L. (1980) *Representation and recognition of the movements of shapes* (Tech Rep. No. 597) Cambridge Massachusetts Institute of Technology, Artificial Intelligence Laboratory

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior *Journal of Personality and Social Psychology, 28,* 28–38

Newtson, D., & Engquist, G (1976) The perceptual organization of ongoing behavior *Journal of Experimental Social Psychology, 12,* 436–450.

Newtson, D., Engquist, G., & Bois, J (1977) The objective basis of behavior units *Journal of Personality and Social Psychology, 35,* 847–862.

Newtson, D., Hairfield, J., Bloomingdale, J., & Cutino, S (1987) The structure of action and interaction *Social Cognition, 5,* 191–237.

Newtson, D., & Rindner, R. (1979) Variation in behavior perception and ability attribution *Journal of Personality and Social Psychology, 37,* 1847–1858.

Parish, D H., & Sperling, G (1987). Object spatial frequency, not retinal spatial frequency, determines identification efficiency *Investigative Ophthalmology and Visual Science, ARVO Supplement, 28,* 359.

Pearson, D E (1981) Visual communication systems for the deaf. *IEEE Transactions on Communications, 29,* 1986–1992

Poizner, H., Battison, R., & Harlan, L. (1979). Cerebral asymmetry for American Sign Language: The effects of moving stimuli. *Brain and Language, 7*, 351–362.

Poizner, H., Bellugi, U., & Lutes-Driscoll, V. (1981). Perception of American Sign Language in dynamic point-light displays. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 430–440.

Riekehof, L. L. (1980). *The joy of signing* Springfield, MO. Gospel Publishing House.

Rindner, R. (1982). *Different units of information in the perception of behavior.* Unpublished doctoral dissertation, University of Virginia, Charlottesville.

Rubin, J. M., & Richards, W. A. (1985). *Boundaries of visual motion* (Tech. Memo No. 835). Cambridge: Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

Sperling, G. (1978). Future prospects in language and communications for the congenitally deaf. In L. Liben (Ed.), *Deaf children: Developmental perspectives* (pp. 103–114). Orlando, FL: Academic Press.

Sperling, G. (1980). Bandwidth requirements for video transmission of American Sign Language and finger spelling. *Science, 210*, 797–799.

Sperling, G. (1981). Video transmission of American Sign Language and finger spelling: Present and projected bandwidth requirements. *IEEE Transactions on Communications, COM-29*, 1993–2002.

Sperling, G., Landy, M. S., Cohen, Y., & Pavel, M. (1985). Intelligible encoding of ASL image sequences at extremely low information rates. *Computer Vision, Graphics, and Image Processing, 31*, 335–391.

Stokoe, W. C., Jr. (1974). Classification and description of sign languages. In T. A. Sebeok (Ed.), *Current trends in linguistics* (pp 345–371). The Hague, The Netherlands: Mouton.

Stokoe, W. C., Jr., Casterline, D., & Croneberg, C. G. (1976) *A dictionary of American Sign Language* Silver Springs, MD: Linstok Press.

Tartter, V. C., & Fischer, S. D. (1982). Perceptual confusions in ASL under normal and reduced (point-light display) conditions. *Perception & Psychophysics, 32*, 327–334.

Thibadeau, R. (1983). *Artificial perception of actions* Unpublished manuscript.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Received October 4, 1988
Revision received June 8, 1989
Accepted June 9, 1989 ■

Journal of Experimental Psychology:
Human Perception and Performance
1990, Vol. 16, No. 2, 445-450

0096-1523/90/$00.75

How to Study the Kinetic Depth Effect Experimentally

George Sperling
New York University

Barbara A. Dosher
Columbia University

Michael S. Landy
New York University

Sperling, Landy, Dosher, and Perkins (1989) proposed an objective 3D shape identification task with 2D artifactual cues removed and with full feedback (FB) to the subjects to measure KDE and to circumvent algorithmically equivalent KDE-alternative computations and artifactual non-KDE processing. (1) The 2D velocity flow-field was necessary and sufficient for true KDE. (2) Only the first-order (Fourier-based) perceptual motion system could solve our task because the second-order (rectifying) system could not simultaneously process more than two locations. (3) To ensure first-order motion processing, KDE tasks must require simultaneous processing at more than two locations. (4) Practice with FB is essential to measure ultimate capacity (aptitude) and, thereby, to enable comparisons with ideal observers. Experiments without FB measure ecological achievement—the ability of subjects to extrapolate their past experience to the current stimuli.

Our article (Sperling, Landy, Dosher, & Perkins, 1989, henceforth, the *source article*) proposed the following. (1) An objective task that involves 53 different shapes to measure shape identification performance in kinetic depth effect (KDE) experiments; (2) an algorithm for the structure-from-motion computations that subjects perform on these and similar stimuli; and (3) a distinction between three kinds of computations. We distinguished (a) the true KDE computation, (b) a KDE-alternative computation—an informational equivalent to the true KDE computation but carried out elsewhere in the brain, and (c) an artifactual computation that arrives at the correct response in a given task but is based on an incidental property of the display. We motivated our discussion by pointing out difficulties in previous work on KDE that we believed could be remedied by measuring objective performance in tasks like ours.

Braunstein and Todd, two experimenters who felt unjustly criticized, wrote a commentary (Braunstein & Todd, 1990) in which they argued that (a) we dismiss legitimate 2D relative velocity cues to KDE as artifactual, (b) our experimental task was not exempt from the criticisms we levied at others, and (c) KDE should be measured in tasks in which subjects are not given feedback about the correctness of their responses (so that the subjects do not learn to use artifactual cues).

Braunstein and Todd's (1990, henceforth, *the critics*) point (a) reflects an unfortunate misreading. In fact, we proposed that "the structure-from-motion algorithm. . .involves finding local 2D velocity minima and maxima and assigning depth values to these locations in consistent proportion to their velocities" (Sperling et al., 1989, p. 839; see also Figure 6, p 838). In this article, we attempt to further clarify the role of relative motion cues in KDE tasks (using 2D dynamic images to answer questions about 3D shape) and in the continuum of mental computations between true KDE and truly artifactual computations. KDE is the perceptual experience of 3D object depth evoked by dynamic 2D images. The opposite end of the continuum is an artifactual computation that arrives at the correct response for the experimental task by using an incidental property of the display. Although artifactual computations need not involve motion cues, in the source article we gave examples of some that do. These computations were called artifactual because motion entered the computation in a way that shortcut the KDE computation. In one example, a measurement of absolute velocity at a single fixed point in the display would have sufficed to yield the correct response. Our aim was not so much to classify computations but to actually deduce the minimal computation that would suffice to solve particular KDE tasks. In a well-constructed task, there is no computational shortcut—the minimal computation is the KDE computation or is essentially equivalent to it.

The critics' arguments (b) and (c) were anticipated and considered in the source article. Here we elaborate the source article's discussion and respond to two newly raised fundamental issues (How should experiments be conducted? and How can a subject's mental computations be exposed, measured, and controlled?) and to other issues (immediacy, practice, and scintillation) that pertain to our specific task.

The 53-Shape KDE Task

Motion Produces an Immediate Experience of KDE

Our task uses 53 different shapes whose surfaces consist of random-dot textures. Each shape is defined by three equally spaced locations. Each location contains either a hill (+1) or

Preparation of this article was supported by the United States Air Force Life Sciences Directorate, Visual Information Processing Program Grant 88-0140.

Correspondence concerning this article should be addressed to George Sperling, Psychology Department and Center for Neural Sciences, New York University, 6 Washington Place, Room 980, New York, New York 10003.

Atch- 12

a valley (−1), or is flat (0). Stimuli are constructed on one of two such sets of locations. A smoothing (spline) transformation merges two hills into a ridge, and merges a three-hill configuration into one spread-out mound. The static random-dot surface has no shape cues and looks utterly flat. When the surface begins to move in a gentle rotary oscillation, it appears instantly to have a particular shape. In the source article, we wrote "All subjects reported that they perceived a 3D surface the first and every time they viewed the high numerosity displays" (p. 830). We have demonstrated these KDE shapes to subjects, to visitors, and to hundreds of observers at numerous lectures and have not yet received even one report of an observer who did not perceive a vivid 3-D shape. We believe this is a natural, ecologically valid test of the shape identification functions that KDE has evolved to perform.

Effects of Practice

Under normal viewing, no learning or practice is needed to perceive these 3-D shapes. The critics' notion that practice is needed to perform the basic task is wrong. However, practice is helpful for some aspects of the task. (a) Subjects must learn to correctly use the naming convention for these shapes. This is usually learned in just a few views of sample stimuli. (b) All subjects remember the shape of the oscillating object, but unpracticed subjects frequently forget the final direction of rotation. With corrective feedback, they learn to report both shape and motion. (c) Some shapes are deliberately made quite similar. For example, in these stimuli, two adjacent hills combine to form an elongated hill. The distinction between a two- and three-hill configuration might be overlooked by a subject who did not received feedback of the correct response, but the distinction is easily learned. However, even highly practiced subjects, with feedback, cannot infallibly discriminate between the two differently oriented three-hill configurations. (d) Small amounts of image degradation are easily tolerated by all subjects. However, to correctly identify shapes when the number of surface dots is grossly reduced or when the signal-to-noise ratio is reduced takes practice.

Feedback or No Feedback?

The critics suggest that KDE experiments be conducted without informing the subject about the correctness or incorrectness of the response (the 3D structure derived from motion), that is, giving the subject no feedback. By eliminating feedback, the critics hope to avoid the problem of the subjects' learning to use incidental cues. We believe that the better way of dealing with incidental cues is to eliminate them, or when that is not feasible, to mask them or render them useless by irrelevant variation (see below). What we address here is the larger question of what can be learned from experiments with and without feedback.

An experiment without feedback is essentially an epidemiological investigation. It investigates the current status of a skill that has been acquired prior to the experimental situation. Therefore, the no-feedback experiment is simplest to interpret when the current test is identical to a previous learning situation. As the experimental test stimuli diverge from the original training stimuli, the experiment must be interpreted in terms of the ease of generalization of previously learned skills to the new testing stimuli. Thus, to test the ability of pilots to discern subtle terrain features in brief glimpses, we would test them with images of natural terrain. Testing pilots with our 53 random-dot shapes, without first affording them an opportunity to practice with feedback, would tell us only how their previously acquired skill generalized; it would not be appropriate for measuring either their previously learned KDE skill or their ability to acquire new skills. Feedback experiments teach us what humans can and cannot learn to do—the limits of human performance. Because experiments with feedback probe the limits of performance, they are the most informative for the discovery of processing mechanisms. In no-feedback experiments, the unknown training situation, and the divergence of test and training, pose problems for theoretical analysis.

Ideal Observers

One kind of investigation that has been particularly informative about human competation is the comparison of human performance with the performance of a statistically ideal observer (Green & Swets, 1966/1974; Sperling & Dosher, 1986). Indeed, the efficiency of information use by humans relative to ideal observers is of practical as well as theoretical interest. Tracing information loss through the stages of sensory analysis yields important insights into sensory processing (e.g., Geisler, 1989; Parish & Sperling, 1987). It would be of great interest to know, in noise-perturbed KDE displays, what the efficiency of human shape identification is in relation to that of an ideal observer. When the efficiency of human perceptual processing is high, we suspect that the task exposes processes that are of evolutionary significance.

To compare the processing efficiency of human and ideal observers, we need to specify exactly what each kind of observer knows about the experimental procedure, such as the a priori knowledge about the probabilities of various kinds of stimuli and the payoffs. This implies an experiment with explicit feedback. To test and evaluate sophisticated models of human mental computation requires us to bring into the laboratory much of the training that often is assumed to have occurred naturally, and it requires more complex and more explicit laboratory procedures than have been used in the past, all with feedback to the subject.

Introspection Versus Objective Measures of Performance

Early psychologists such as Wundt (1905) and James (1890) attempted to distinguish the new discipline of experimental psychology from the natural sciences by emphasizing different methodologies. They were especially concerned with how things appeared to them—introspection—rather than with measurable skills and abilities. An important component of the development of psychology has been the move away from introspection—now viewed as an extended verbal report—and toward behaviors that are simpler, more easily measured,

and more directly related to evolutionary development. Behavioral investigations of KDE tend to require more work than the corresponding introspective investigations, but that is a tax to which perceptionists have become accustomed. From a formal point of view, the subject's report in an introspective procedure "tell me what you see" and in an experiment without feedback "is what you see a circle or a cylinder?" have in common that they report the subjective appearance of the world unconstrained by feedback of objective reality, and they differ primarily in the degree of constraint on the response. In studying KDE, which is a critical component of a structure-from-motion system that has evolved to meet a biological demand, it is important not to stop at the point of studying appearances (perceptions) but to continue to investigate how these perceptions ultimately govern performance.

KDE, Alternative Computations, and Artifactual Computations

Motion Input to KDE Computation

A stationary surface covered with random dots appears quite flat. As soon as the surface starts to rotate, it is perceived as having depth—the kinetic depth effect (KDE). It is useful to think of KDE in relation to stereopsis, which is the perception of depth induced by the disparity differences between images in the left and right eyes. In perfect analogy to stereopsis, the disparity differences between dots in two successive frames of our random-dot displays suffice to give subjects a good impression of KDE depth (Landy, Dosher, Sperling, & Perkins, 1983; Landy, Sperling, Perkins, & Dosher, 1987; Todd, 1988). Because only velocity is computable from two frames, and not acceleration (which requires three), KDE perceived in two-frame displays implies that acceleration is not needed for KDE. Similarly, constructing a sequence of frames so that each individual dot has a lifetime of only two frames yields high shape identification accuracy (Dosher, Landy, & Sperling, 1989). These, and other of our results, imply that a velocity flow-field is a sufficient stimulus for KDE. The identities of the moving dots are preserved only long enough to yield a velocity estimate; subsequently, only the velocities and not the dots themselves are needed for the KDE structure-from-motion computation.

The human perceptual system uses two fundamentally different computations to extract motion flow-fields. First-order motion analysis is served by motion detectors that approximate a spatiotemporal Fourier analysis based on stimulus contrast (Adelson & Bergen, 1985; van Santen & Sperling, 1984, 1985; Watson & Ahumada, 1985). Second-order (nonFourier) motion analysis uses more complex stimulus properties and invariably involves rectification (e.g., the absolute value of contrast; Chubb & Sperling, 1988, 1989). Dosher et al. (1989) showed that the first-order system was the predominant contributor to KDE in our random-dot stimuli (because only it had sufficient spatial resolution), although second-order motion computations could yield limited KDE under special conditions.

Structure-From-Motion Computation

Given that the first-order velocity flow-field is the input stimulus for KDE, we specify a computation that would be sufficient to extract 3D shape from our 2D stimuli. Considering that the stimuli were viewed by parallel projection and with the axis of rotation perpendicular to the viewing axis, the principle is the same as for motion parallax: The difference between local object depth and the mean object depth is proportional to the difference between the local image velocity and the overall image velocity. The sign of the proportionality (+ or −) is undetermined. To discriminate among our 53 stimuli, it is not necessary to compute depth everywhere in the image. For example, it suffices to determine depth (velocity) at the six locations in which a 1, 0, or −1 could be placed during stimulus construction. Alternatively, it would be sufficient to locate the velocity minimum or maximum, to derive a three-valued descriptor of the extremum, and to interpolate smoothly to the reference plane everywhere else. (A plausible peak descriptor would take the values *normal* [single peak used in construction], *elongated* [the combination of two peaks], or *enlarged* [the combination of three peaks].) Although the precise location of the extremum would (in the absence of noise) be sufficient for an ideal detector to discriminate between the 53 shapes, human observers use the actual shape in the neighborhood of the extremum in their judgment.

KDE Versus Non-KDE Computations

In the source article, we proposed a continuum of computations ranging from a true KDE computation, in which 3D shape assignment is accompanied by the introspective impression of depth structure, to artifactual computations. Artifactual computations can be eliminated by appropriate stimulus manipulations. The more troublesome possibility is a *KDE-alternative computation* that is algorithmically equivalent to the true KDE computation and may share the same motion inputs, but is not accompanied by the introspective impression of a shape in depth. In the source article, we provided a task in which the subject viewed six isolated patches in which dots moved at the velocity of six key locations in our KDE displays. Performance in this task, which did not involve KDE, demonstrated not only that an alternative computation could occur, but that the pattern of responses and errors produced by the alternative computation mirrored the response pattern in the KDE task.

Here we consider perhaps the most troublesome possibility: true KDE supplemented by other computations. Suppose, for example, in viewing one of our complex random-dot shapes, the subject (a) experiences weak KDE and sees a hill and an ambiguous area, (b) observes that dots in the hill and in the ambiguous area of the display are moving in the opposite directions, (c) infers that these two subareas represent the opposite depth planes, and (d) uses both sources of information in his response. How can one deal with the problem of discovering the algorithms underlying KDE performance in KDE tasks and their precise implementation? The critics assume that subjects naturally use KDE in KDE tasks, and that by not giving subjects feedback on the correctness of their

responses, subjects will not learn to use artifactual or alternative computations.

Because subjects continuously practice natural KDE tasks in everyday life, which provides feedback, they may normally use conscious or unconscious non-KDE computations to help them to derive structure from motion. Non-KDE computations are not necessarily restricted to the laboratory, and it is worthwhile to develop methods to measure them. Because KDE is distinguished from a KDE alternative by a subjective impression of perceived depth, subjective reporting is a necessary component in the perceptionist's arsenal. By clever experimentation, artifactual computations usually can be distinguished from KDE and KDE alternatives. Whereas the critics rely primarily on subjective reports and introspection in approaching the problem of artifactual and alternative computations, we feel it is necessary to augment the introspective approach with successive refinements in experimental procedures to gain control over critical factors as they are discovered. For example, in our shape identification task, using shapes of considerable complexity makes artifactual computations (which rely on small amounts of incidental information) almost useless. Keeping viewing times fairly brief create a relative disadvantage for serial (versus parallel) computations. To further discriminate between KDE and non-KDE computations, the source article proposed dual tasks that selectively interfered with the mental processing resources required for alternative or artifactual computations.

Multilocation Motion Tasks

The source article argued, and the critics appear to agree, that the 53-shape identification task is less susceptible to artifactual computations than earlier tasks. However, the critics argue that the distinction between the 53-shape task and previous KDE tasks is not based on any fundamental principle, but is based merely on the number of locations from which velocity information must be extracted in order to perform the task. In rebuttal, we show here that the distinction between one or two versus six locations is critical (because either the first- or second-order motion system can provide simultaneous velocity information about one or two locations, but only the first-order system can provide information about six). In the source article, we showed that the structure of the earlier tasks permitted information from only one or two locations to discriminate perfectly between alternative responses when the same information would have been insufficient to construct a 3D shape representation.

Dosher et al. (1989) showed that complex shape identification, based on motion at three or more locations, operates very differently from motion extraction at one or two locations. They used stimuli that were designed to selectively stimulate either the second-order motion system or both the first- and second-order systems and compared them in four tasks: the 53-shape identification task, a threshold detection task for motion in a single patch, a threshold direction-of-motion task in a single patch, and a motion segmentation task that required finding the one-of-nine possible locations at which there was an odd direction of motion. Manipulations that disrupted first-order motion information (such as rapidly alternating the contrast of stimulus dots on a gray ground between black and white) reduced identification of 3D shapes to near chance levels. Such manipulations also severely impaired the motion segmentation task. Performance was equivalent to sophisticated guessing that uses velocity information based on only one or two locations. In contrast, detection performance and direction-of-motion judgments in single patches of planar motion survived disruption of first-order information.

There are several reasons why artifactual computations based on one or two locations may survive first-order motion disruption when KDE computations cannot. Velocity information about one or two locations may be obtained either by tracking individual image features, or by information processed through a second-order system. However, second-order motion computations are based on some form of rectification (either half wave or full wave) that implies loss of information in relation to first-order motion (Chubb & Sperling, 1988; Sperling, 1989). Empirically, second-order information is virtually restricted to the fovea, and even there it is of low spatial resolution (Chubb & Sperling, 1988, 1989). Not surprisingly, when the recovery of 3D shape requires simultaneous information about motion in three or more locations, it is extremely vulnerable to disruption of first-order information.

The Dosher et al. (1989) results are one cogent empirical example of why we make the distinction between tasks that require observation of only one or two locations and those that require more. The 53-shape KDE task has been demonstrated to require the first-order motion system. Indeed, first-order motion appears to be the essential input to all complex KDE discriminations. Tasks that can be solved by extracting velocities at two locations do not necessarily involve either KDE or the first-order motion system. Our criticism of other tasks has been that they easily can be solved by computations that are less than the whole KDE computation. Indeed, it would be a step forward if the KDE experimenters offered a plausible KDE computation against which performance in their tasks could be measured.

How to Deal with Artifactual Cues

Dot Density: An Artifactual Cue in KDE Experiments

The source article investigated structure from motion. In this context, a nonmotion shape cue such as a local variation in 2D dot density is artifactual. (Of course, in a shape-from-texture task, texture-density cues would be primary.) We eliminated static (single-frame) density cues by resampling of stimulus dots when necessary to maintain a uniform 2D density on each successive frame of the motion stimulus, as follows. The stimulus field was divided into 100 fixed areas of equal size. Each area was constrained to have three dots in every frame. The 3D motion of the surface between frames causes dots to wander in and out of areas, some areas having a net inflow and others a net outflow. Therefore, to satisfy the constraint of having a constant number of dots, dots were added or subtracted at randomly chosen locations within local areas. The fraction of new-plus-discontinued dots divided by the fixed number of dots in an area is the scintillation fraction. Our displays typically required 5% frame-to-frame scintillation.

In the source article, we measured the density artifact by producing a display that had no motion cue, only the extracted density cue. With the density cue alone, only 1 of our 3 subjects achieved above-chance shape identification. On the other hand, with the motion cue alone (without the density artifact), shape identification was almost as good as with motion and density cues together. The slight impairment was probably due to the scintillation that accompanied removal of the density artifact.

Scintillation

The critics point out that our method of eliminating the density cue introduces a region-specific variation in scintillation, which is itself a possible cue in a shape identification test. They prescribe avoiding practice and eliminating feedback as the remedy. Are restricting the opportunity to practice and eliminating feedback the optimal methods of dealing with artifactual cues? Whatever the degree of practice or feedback, we believe that it is better either to determine the possible effectiveness of significant artifacts or to eliminate them, as we did the density cue. Simply assuming that lack of practice will suffice is not adequate. Indeed, the possibility that scintillation variations might be a shape cue was considered in the source article; it was dismissed because scintillation is an even weaker cue than the extremely weak density cue. That is, when displays are constructed without motion cues but with only a density cue or a scintillation cue, it obviously is harder to perceive areas where scintillation differs from the average than where density differs from the average (Lappin, Doner, & Kottas, 1980). This is because random error in dot density as a direct cue to 3D slope occurs only because of sampling error (limited number of dots) and the fineness of the 182 × 182 pixel grid, whereas random error in the relation of scintillation to a frame-to-frame change in 3D slope occurs because of the coarseness of the 10 × 10 grid of local areas within which density was kept constant. That is, because of the way stimuli were constructed, scintillation density was an objectively less reliable cue to shape than was dot density. Like the dot-density cue, the scintillation cue in our displays can be measured alone and it can be compensated or masked.

Displays were constructed that had a pure scintillation cue, without the changing-density or motion cues. The only subject who was able to perform above chance with the isolated density cue also viewed the new displays. With a pure scintillation cue, it was clear that his performance in a shape identification task would have been even lower than that with a density cue, although we did not feel it was worthwhile to run a formal experiment. Conversely, displays were constructed with normal density-controlled KDE cues, but with extraneous scintillation added uniformly throughout the display to mask the scintillation cue. Shape identification in these displays (with the scintillation cue rendered ineffective) appeared essentially equivalent to normal KDE displays. However, adding extraneous scintillation reduces the signal-to-noise in the stimulus, and more extended observations undoubtedly would reveal a slight impairment—not due to the loss of the scintillation cue but to the added scintillation. The bottom line for displays that are not scintillation-corrected is that the residual scintillation cue could be used to make some extremely coarse discriminations (e.g., there probably is more scintillation on the left than on the right of the display) that may support above-chance shape identification for extremely sophisticated subjects; it is not a significant factor when normal motion cues are available.

General Procedures for Dealing with Artifactual Cues

The procedures used to measure and to eliminate the dot-density and scintillation cues illustrate general principles. If a particular artifact yields above-chance guessing, (a) measure the strength of the artifactual cue in isolation and (b) construct displays in which the cue is eliminated, masked, or rendered useless by irrelevant variation. Creating the cue in isolation is useful because bounds on the possible strength of the cue can then be determined. For example, dot density and scintillation were extremely weak cues. Eliminating artifactual cues in the displays of interest is an ideal solution, but is not always possible. Thus, dot-density cues could be eliminated, but this introduced scintillation cues that could not be eliminated but could be masked by adding still more scintillation. It was not necessary to use the third general method of dealing with unwanted cues—introducing irrelevant variation. For example, irrelevant variation is used to eliminate motion extent as an artifact in velocity estimation (McKee, Silverman, & Nakayama, 1986).

In our experience, it has never been necessary or preferable to deal with possible artifactual cues by using naive subjects without feedback and hoping that the subjects do not use the artifactual cues. To review our previous discussion: the problem is that, for optimal performance, subjects must also learn to optimally use the relevant cues, and this requires practice with feedback.

Summary and Conclusions

1. The extraction of 2D relative velocity is a basic substrate for deriving 3D structure from dynamic visual stimuli for both the true KDE or KDE-alternative computations.

2. The 53-shape lexicon for our identification task presents an ecologically valid test of shape recovery (KDE) for complex depth surfaces.

3. Practice in the 53-shape task serves to optimize identification performance; however, practice is not necessary to immediately perceive vivid KDE.

4. Properly conducted, experiments with feedback can measure the limits of human capacity; experiments without feedback measure the ability of subjects to generalize from their past experience to the experimental stimuli.

5. Excluding feedback in KDE experiments does not eliminate the possibility that artifactual cues may generate a correct response, it merely confuses the issue.

6. Scintillation is an insignificant cue in the 53-shape stimuli.

7. Deriving high-resolution 3D structures from 2D dynamic displays requires the first-order motion processing system. In moving-dot displays such as ours, the second-order motion system cannot be used to solve tasks that require simultaneous access to velocities at more than two locations. Therefore, to isolate the KDE performance supported by first-order motion

processing, it is desirable to use stimuli that require computation of velocities at more than two locations.

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A, 2,* 284–299.

Braunstein, M., & Todd, J. (1990). On the distinction between artifacts and information. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 211–216.

Chubb, C., & Sperling, G. (1988). Drift-balanced random stimuli: A general basis for studying non-Fourier motion perception. *Journal of the Optical Society of America A, 5,* 1986–2007.

Chubb, C., & Sperling, G. (1989). Two motion perception mechanisms revealed through distance driven reversal of apparent motion. *Proceedings of the National Academy of Sciences USA, 86,* 2985–2989.

Dosher, B. A., Landy, M. S., & Sperling, G. (1989). Kinetic depth effect and optic flow: I. 3D shape from Fourier motion. *Vision Research, 29,* 1789–1813.

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review, 96,* 267–314.

Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics.* New York: Krieger. (Original work published 1966)

James, W. (1890). *Principles of psychology.* New York: Holt.

Landy, M. S., Dosher, B. A., Sperling, G., & Perkins, M. E. (1988). The kinetic depth effect and optic flow: II. First- and second-order motion. *Mathematical Studies in Perception and Cognition* (Tech. Mem. No. 88-4). New York: New York University.

Landy, M. S., Sperling, G., Perkins, M. E., & Dosher, B. A. (1987). Perception of complex shape from optic flow. *Journal of the Optical Society of America A, 4,* 108.

Lappin, J. S., Doner, J. F., & Kottas, B. L. (1980). Minimal conditions for the visual detection of structure and motion in three dimensions. *Science, 209,* 717–719.

McKee, S. P., Silverman, G. H., & Nakayama, K. (1986). Precise velocity discrimination despite random variations in temporal frequency and contrast. *Vision Research, 26,* 609–619.

Parish, D. H., & Sperling, G. (1987). Object spatial frequency, not retinal spatial frequency, determines identification efficiency. *Investigative Ophthalmology and Visual Science, ARVO Supplement, 28,* 359.

Sperling, G. (1989). Three stages and two systems of visual processing. *Spatial Vision, 4,* 183–207.

Sperling, G., & Dosher, B. (1986). Strategy and optimization in human information processing. In K. Boff, L. Kaufman, & J. Thomas (Eds.), *Handbook of perception and performance* (pp. 2-1–2-65). New York: Wiley.

Sperling, G., Landy, M., Dosher, B., & Perkins, M. (1989). Kinetic depth effect and identification of shape. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 826–840.

Todd, J. T. (1988). Perceived 3D structure from 2-frame apparent motion. *Investigative Ophthalmology and Visual Science, Supplement, 29,* 265.

van Santen, J. P. H., & Sperling, G. (1984). Temporal covariance model of human motion perception. *Journal of the Optical Society of America A, 1,* 451–473.

van Santen, J. P. H., & Sperling, G. (1985). Elaborated Reichardt detectors. *Journal of the Optical Society of America A, 1,* 300–321.

Watson, A. B., & Ahumada, A. J., Jr. (1985). Model of human visual-motion sensing. *Journal of the Optical Society of America A, 1,* 322–342.

Wundt, W. (1905). *Grundriss der Psychologie* [Foundations of psychology.]. Leipzig, Germany: Englemann.

Received August 28, 1989
Revision received October 12, 1989
Accepted October 13, 1989 ■

Publication Practices and Scientific Conduct

The recent disclosures of fraud in the conduct of research, reporting of research, or both in a number of scientific disciplines have prompted a widespread program of self-examination of publication practices and ethics.

The editor joins with APA in reminding authors of the principles of good publication practices and scientific conduct. Prospective authors are directed to the *Publication Manual of the American Psychological Association* (3rd ed.) and to the "Instructions to Authors" printed in this issue. The requirements of data availability, replicability, authorship credit, ethical treatment of subjects, and primary publication of data are important—they are meant to ensure responsible science and appropriate use of scarce and valuable resources.

George Sperling,
Comparison of Perception in the Moving and Stationary Eye.
In E. Kowler (Ed.), *Eye Movements and their Role in Visual and Cognitive Processes.*
Amsterdam, The Netherlands: Elsevier Biomedical Press,
1990. Pp. 307-351.

Atch-13

Reviews of Oculomotor Research
Volume 4

Eye Movements and Their Role in Visual and Cognitive Processes

Edited by

Eileen Kowler

Department of Psychology
Rutgers University
New Brunswick, New Jersey
U.S.A.

1990

ELSEVIER

AMSTERDAM · NEW YORK · OXFORD

Eye Movements and their Role in Visual and Cognitive Processes
Edited by Eileen Kowler

Table of Contents

Chapter 1, Pp. 1-70

The Role of Visual and Cognitive Processes in the Control of Eye Movement
by Eileen Kowler

Chapter 2, Pp. 71-114
Predictive Control of Eye Movement
by M. Pavel

Chapter 3, Pp. 115-212
The Role of Eye Movement in the Detection of Contrast and Spatial Detail
by Robert M. Steinman and John Z. Levinson

Chapter 4, Pp. 213-261
Binocular Eye Movements and the Perception of Depth
by Han Collewijn and Casper J. Erkelens

Chapter 5, Pp. 263-287
Eye Movement and Visual Localization of Objects in Space
by Alexander A. Skavenski

Chapter 6, Pp. 289-305
The Role of Eye Movements in the Perception of Motion and Shape
by Hans Wallach

Chapter 7, Pp. 307-351
Comparison of Perception in the Moving and Stationary Eye
by George Sperling

Chapter 8, Pp. 353-393
Eye Movements in Visual Search: Cognitive, Perceptual and Motor Control Aspects
by Paolo Viviani

Chapter 9, Pp. 395-453
Eye Movements and Reading
by J. Kevin O'Regan

Chapter 10, Pp. 455-485
Eye-Movement Models for Arithmetic and Reading Performance
by Patrick Suppes

Eye movements and their role in visual and cognitive processes
E. Kowler, Editor

CHAPTER 7

Comparison of perception in the moving and stationary eye

George Sperling

Human Information Processing Laboratory, Department of Psychology and Center for Neural Sciences, New York University, New York, NY 10003, U.S.A.

1. Introduction

There are many different kinds of eye movement, each of which sweeps the image of the world across the retina in its distinctive way. Each kind of eye movement serves a unique function and in turn is served by a unique perceptual-motor process. Undoubtedly, the visual system has evolved special adaptations for this mode of operation. This chapter probes the question of to what extent these information-processing adaptations operate independently of the eye movements themselves.

Simulated eye movements: imposed motion. An important tool is the simulated eye movement: an image sequence which produces on a stationary retina precisely the same image motion that the moving eye would have produced. Comparing perception when image motion is produced by eye movements with perception under imposed motion offers substantial insights into both the mechanisms and the purposes of eye movements. To provide a frame of reference and to illustrate the general similarities between processing principles, all kinds of eye movements are considered briefly. The primary emphasis is on saccades because this is the domain the author has studied experimentally.

Restriction to eye-movement-induced image motion Eye movements are usually measured in the laboratory with the body at rest and the head held stabilized by a 'biteboard' – an impression of the teeth of the observer which holds the observer's teeth, and thereby the head, in place. Of course, in normal viewing, the head and body move freely. The motion of images across the retina caused by uncompensated movements of the head and body is large compared to the image motion induced by some of the eye movements considered here. The nature and consequences of image motion produced by eye, head and body movements are considered in Ch. 3 of this volume (Steinman and Levinson). In the present chapter, only image motion caused by eye movements, principally saccadic eye movements, with the head and body held stationary is considered and compared with image motion produced by corresponding object movements. The descriptions of the eye movements themselves are very brief here because they are taken up in great detail in other chapters of this book.

2. Tremor

High-frequency tremor is a generic term applied to eye movements that are typically about 30–70 Hz, with amplitudes less than the width of a single cone, approx 0.5 min (Ratliff and Riggs, 1950; Yarbus, 1957). The blur circle under normal vision is about 2 min or more (Krauskopf, 1962; Westheimer and

Campbell, 1962), so the effects of tremor on the retinal image would be obscured by the much greater blur produced by the eye's optical imperfections. High-frequency tremor is probably irrelevant for normal vision. Occasional attempts to attribute some useful purpose to tremor, such as Yellott's (1987) proposal that tremor may act to smooth over irregularities in receptor spacing, have the aforementioned problem that the tremor is negligible compared to the much larger optical aberration.* Indeed, Krauskopf (1957, 1960) imposed tremor frequencies on retinally stabilized images and found that they did not improve the visibility of his test stimulus (a single line), in some instances imposed high-frequency motions were detrimental. Based on these measurements, I would predict that if normal tremor could suddenly be removed from vision, the change could probably not be noticed even in the fovea under psychophysical viewing conditions. Elsewhere in the visual field, or in the fovea under natural viewing with unrestrained head and body movements adding to retinal image motion, high-frequency tremor would seem to be far below the threshold of visibility.

3. Slow control movements

Slow control refers to the involuntary smooth motions of the eye that occur during steady fixation on a stationary target. A typical velocity for slow control movements, when the head is stabilized, is about 5–10 arc min per s (Ratliff and Riggs, 1950; Steinman et al., 1973), and a typical oscillatory frequency is about 2–3 Hz. Slow control movements may or may not be interrupted by small saccades, and the saccades may move the eye nearer or further from its intended fixation (see Ch. 1 of this volume for a more detailed review).

Are slow drift movements important for vision? When slow control movements and saccades are removed in stabilized vision, the image fades within several seconds. Visibility is restored to a stabilized image by imposing slow image oscillations. For optimal restoration of visibility, velocities higher than those typically observed in biteboard viewing are required (see Ch. 3 of this volume).

There are two issues here: (1) the role of drift image-motion in visibility, and (2) the appearance of movement in drifting images. I will consider first the appearance of movement in drifting images. Suppose that image drift were due to imperfect oculomotor control. Then, retinal drift would not be accompanied by signals of intentional motor movement; it would be residual image instability that was uncompensated by motor movement. Suppose also that proprioceptive information is unimportant when the eye is maintaining stable fixation. Then, if image drift were to be recorded from a subject and later produced as an imposed signal on a stabilized retina, the subject would have no way of discriminating the original drift due to image instability from the imposed motion. That is, image drift resulting from resolution failures of the oculomotor system logically cannot be discriminated from the same image drift imposed on a stabilized retina. Imposed drift will be discriminably different from natural drift when the amplitude of the imposed drift is artificially increased to be greater than natural drift. Just how much increase of artificial over natural drift is necessary for discrimination is not known. And we do not know to what extent typical drift movements, natural or imposed, produce apparent motion. This problem needs experimental study.

4. Slow drift with head free-moving

Not all head movements are intentional. Even when trying to hold the head as still as possible, as for example in walking, the head inevitably oscillates. The best possible compensation for retinal image slip with small head movements leaves a residual, uncompensated image slip of about 0.5 deg per s (Skavenski et al., 1979), with modest voluntary head rotation the image slip increases 10-fold to 5 deg per s (Steinman and Collewijn, 1980; Steinman and Levinson, Ch. 3 of this volume). In carefully controlled experimental situations, when motion of the head is induced by rotating the subject in a chair, the eye movements are invariably imperfect, and the proportion of compensation for head movement depends on the amplitude and frequency of the head movement (Skavenski et al., 1979). These authors found that the gain of compensatory movement decreased as the amplitude of oscillation decreased, and suggested an automatic adjustment of gain sufficient to maintain clear vision.

Despite the large image slip as the head moves around, the world does not appear to move, except in extreme situations. Remarkably, acuity for gratings hardly suffers: there is a slight decline in contrast sensitivity at high spatial frequencies and a slight improvement at spatial frequencies below 5 cycles/deg (Steinman et al., 1985).

When an image on the eye is stabilized, and image motions which mimic the large motions of normal unstabilized viewing are then imposed on this image, what does it look like? And how does the image movement affect acuity? Again, this experiment has not been carried out, although it seems obvious that the image would appear to move around. The effect of imposed motion on acuity has been studied only with very simple procedures. Kelly (1979) tested contrast sensitivity for sine gratings in stabilized vision with imposed constant image velocities, and found large acuity changes dependent on imposed velocity: a great decline in acuity at high frequency of imposed image movement, a small improvement at low frequencies. Steinman et al. (1985) tested acuity under comparable conditions of natural image motion in the same range of velocities and found smaller acuity losses. In Steinman et al.'s procedure, the image moved back and forth as the subjects moved their heads. In Kelly's procedure, the image moved at a constant velocity for several seconds before the subject responded. If Kelly's results with prolonged viewing of constant image velocities generalized to back-and-forth image motions, then we would have to conclude that imposed image motion is harmful to grating acuity, over a range of image velocities that does not lead to acuity loss in natural viewing (see Ch. 3).

5. Smooth pursuit

Smooth pursuit refers to the smooth following of external image movement. Whether there is a significant distinction between the oculomotor subsystem responsible for following external image motion and the oculomotor subsystem that follows intrinsically caused image motion (produced, for example, by head or body movements) has been debated for many years. The velocity of smooth pursuit can be remarkably high, up to or exceeding 100 deg/s for large-amplitude motions (Collewijn et al., 1985; Meyer et al., 1985). The gain of smooth pursuit depends on many factors: notably the pattern of motion (waveform, frequency, amplitude, velocity of constant velocity motion, etc.) and on past history. The luminance contrast of the moving stimulus pattern seems not to be critical (Haegerstrom-Portnoy and Brown, 1979; Winterson and Steinman, 1978).

5.1. Localization in smooth pursuit

Flashed targets. A flashed target seen during smooth pursuit is very accurately localized relative to the body (Hansen, 1979). Untracked objects seen during pursuit are also accurately localized. Taken together, Hansen's results establish that eye position is accurately known during smooth pursuit, and the knowledge of eye position is available to perceptual judgements relative to body position. However, the extent to which this extra-retinal knowledge is available to visual judgements of the relative positions of targets flashed during smooth pursuit is less clear (see Ch. 5 of this volume). The extent to which flash localization during real pursuit differs from localization during simulated pursuit is not known (cf. sections on spatial localization during saccades and simulated saccades.)

* The suggestion that small eye movements might somehow overcome image perturbations induced by receptor irregularity, i.e. by higher visual centers assuming that the retinal receptors are embedded in a regular grid when they are not, encounters, additionally, a profound logical problem because receptors are typically as informative when they are spaced irregularly as regularly (Maloney, 1988, 1990; Yellott 1987). The role of eye movements in correcting perception at a larger scale is considered later.

Reconstruction of the trajectories of moving objects. Suppose that a moving dot describes a perfect circle. When the eye is fixated, the circular trajectory is perceived veridically (correctly). During linear smooth pursuit of another target, the retinal projection of the circle becomes distorted. The perceptual system for computing localization of the eye relative to the body only partially solves this distortion problem. That is, the perceived path is a composite of the distorted retinal path and the true circular path in external space. (See the review by Mack, 1986, and Chs. 5 and 6 for further discussions of these illusions.) The issue of how spatiotemporal relations between stimuli that occur during eye movements might be reconstructed will be reconsidered in the section on saccades.

5 2 *Simulated pursuit movements*

Acuity. Brian Murphy (1978) produced, in stationary eyes, image velocities that were the same as the image slip velocities previously measured during pursuit movements. He measured contrast sensitivity for a 5 cycles/deg grating in both conditions and found no difference. Velocities above 2 deg/s produced equivalent acuity losses in both viewing conditions Kelly's (1979) data cited above appear also to be related here, and to yield a conflicting conclusion.

5 3. *Attention during smooth pursuit: search task*

From many experiments (e g. Dodge and Fox, 1928, Dubois and Collewijn, 1979; Kowler et al., 1984) we know that subjects can selectively choose which one of several retinal stimuli to track. Even when retinal position is controlled, the tracking instruction determines what is tracked. The question Khurana and Kowler (1987) sought to answer was: can a subject track one stimulus while attending to another?

The subjects in Khurana and Kowler's (1987) experiments tracked a moving 4×4 letter array with smooth-pursuit eye movements. During the movement interval, all the characters of the array were

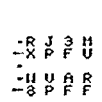


Fig 1 Example of a display used by Khurana and Kowler (1987) Four rows of characters are visible at once. The arrows indicate the relative speeds of motion to the left. The subject is instructed to track an (invisible) point between the rows at the speed of one of the rows. The array contains two numerals among the letters. one in a fast row and one in a slow row. The subject's task is to report both numerals.

briefly changed, with two of the former letters being replaced by numerals; the subjects' task was to detect these target numerals. In various procedures, the odd rows (1 and 3) were moved at twice or at 1/2 the velocity of the even rows (Fig. 1) Two targets occurred one target in the even rows, another in the odd rows The subject was instructed to report both targets on every trial. On different trials, the subject was instructed to smoothly pusue either the even or the odd rows, always with fixation in the middle of the array.

The results of this procedure showed that the subject reported the letters in the tracked rows better than the untracked rows regardless of the actual retinal slip (because the retinal velocities were too low to degrade performance) That is, visual attention was naturally linked to the pursuit attempt.

In a control experiment, tracking and attention instructions were manipulated separately. Subjects were asked to smoothly pursue one set of rows while attending to the other. All the conditions taken together provide a cross-design in which the positively correlated factors of the main experimental condition (row-to-be-attended and row-to-be-tracked) are negatively correlated in the control condition Thus, we can estimate the effect of the two separable independent variables (tracking instruction, attention instruction) upon the two dependent variables (tracking accuracy, search accuracy)

The results showed that tracking of pursued but unattended rows was slightly worse than tracking of these same rows when they were attended (attention instructions only slightly affect tracking). Search performance on the attended but untracked rows was slightly better than search in the unattended untracked rows, but never approached the performance on tracked, unattended rows, again, regardless of retinal slip (attention instructions only slightly affect search). In other words, the tracking instruction, not the attention instruction, controls the accuracy of visual search.

Attention is inextricably linked to tracking The following picture emerges of the role of attention in tracking: tracking instructions have big effects, attention instructions have only small additional effects. For a given rate of retinal slip, the only way that tracking per se can influence search accuracy is indirectly through attention. That an attention-instruction has almost no effect independently of tracking-instruction implies that attention is inextricably linked to smooth pursuit.* Tracking a row depends on attentional selection and only an insignificant attentional residue remains to be assigned elsewhere.

In the imposed-motion control procedure for Khurana and Kowler's experiment, the eye remains stationary as the letter rows drift across the field. The subject's attempt to fixate a stationary point while attending to one pair of moving rows leads either to a loss of fixation as the eye drifts in the movement direction or to a loss of search accuracy in the moving rows. Thus, the data obtained with eye movements and imposed movements are essentially equivalent. It follows that the attentional resources needed to maintain the eye fixated on a stationary point are not essentially different from those needed to maintain smooth pursuit at the image velocities Khurana and Kowler used. This result lends support to the view that fixation and smooth pursuit are governed by the same mechanisms (Nachmias, 1961).

* To have reached the conclusion that subjects *could not* dissociate attention from tracking (versus simply that they *habitually did not* dissociate attention from tracking) it was essential that Khurana and Kowler's subjects actually received training – feedback as to the correctness of the responses – and despite the training failed to learn to perform both tasks.

6. Saccades

Saccades are voluntary, quick, ballistic eye movements that take the eye from one fixation point to the next. Saccades range in extent from several minutes of arc to more than 70 degrees. A typical saccade of 4 degrees is well described as a ramp function of time in which the eye travels from its initial to its final position in about 15–20 ms, and then remains relatively still in the final position for 200 ms or longer. During active search of a display or scene, saccades may occur at a rate of 4 per second, but slower saccade rates are more typical. The effect of saccades is to convert the input to the visual system into a sequence of up to 3 or 4 relatively stationary images per second, with rapid transitions between the images.

There are many provocative issues concerning saccades. How is it that the world seems to remain stationary during a saccade even though the saccade-like image sequence would be perceived as a vigorous jump if it were imposed on a stationary retina. The visual image of the world is smeared across the retina during saccades yet, on the whole, we are unaware of seeing such a smear. Is this because visual sensitivity is reduced during saccades? Are there special perceptual mechanisms designed to utilize and link information acquired during successive saccades? For example, do saccades initiate visual processing episodes? And are relative spatial coordinates defined by saccadic eye movements or by head movements inherently more useful than coordinates defined equally accurately by image movements or by other more indirect means? Is attention inextricably linked to saccades (as it is to smooth movements) or can attention move oppositely to a saccade? In what sense are saccades an optimal solution to the ecological problem confronting a visual system?

Some of these issues have been with us almost since saccades were first described by Javal in 1878. For example, with respect to the issue of vision

during saccades, an early demonstration is due to Woodworth (1906). He executed a saccade from one side of a rotating wheel to the other and observed that he was able to see clearly the spokes that happened to be traveling at the same rate as his eye. Thereby Woodworth demonstrated his capacity for sharp vision during a saccade. After a review of similar experiments in his textbook Experimental Psychology, he wrote *"given the same retinal stimulation it makes no difference whether it is the eyes or the external field that moves"* (Woodworth, 1938, p 593, italics in the original) Woodworth's conclusion was so intuitively unsatisfactory that the issue of visual suppression during saccades has been reexamined many times, and we examine it once again

6 1 *Is vision turned off during saccades?*

The rumor that vision is dead during saccades is grossly exaggerated Some time ago, eager to observe such an effect, I put together an apparatus for studies of vision during saccades A gas discharge lamp illuminated a thin slit before, during, or just after saccades. The illumination flash was very brief, contained within 40 microseconds, so there would be no significant retinal smear of the slit during a saccade Before conducting formal experiments, I wanted to check whether the apparatus actually triggered flashes during saccades I quickly discovered that, in viewing the slit against a dark background, it was not possible to ascertain whether a flash had occurred during a saccade simply by noting the appearance of the flash A flash that occurred during a saccade did not look any different from a flash that occurred before or after. Obviously, if there were a change in visual sensitivity during saccades, it was too small and too subtle to make an obvious difference in the appearance of a suprathreshold flash. Nevertheless, there have been many reports of raised thresholds for stimuli presented during saccades (for a review, see Volkmann, 1986) How can this be?

The question of whether visibility is altered during a saccadic movement must be resolved by the proper control experiments. In the control experiment, it is essential to produce on the stationary eye precisely the same sequence of stimuli that the saccading eye produces for itself when its threshold is tested during saccades There are two reasons for this requirement. First, the movement of fixation points and other fixed stimuli across the retina during a saccade can affect the threshold for a dim test field, and the amount of this visual masking must be measured in the control experiment. Second, the apparent location of a stimulus flashed during, or slightly before or after a saccade, will not generally correspond precisely to its objective location. From the observer's point of view, there are locational *uncertainties*, indeed, even locational illusions, in which the test flash appears to have occurred at an unexpected location. Because, as is demonstrated below, there are many similarities in spatial localization during saccades and during the equivalent imposed image motion, control experiments may provide reasonable estimates for locational uncertainty and locational illusions as well as for incidental masking. We defer the issue of altered visibility during saccades until after considering localization

6 2 *Spatial localization during saccades*

There are now many studies of the localization (in which observers indicate the apparent location) of test stimuli that are flashed briefly during saccades I describe studies (Sperling and Speelman, 1964, 1965, Sperling, 1966) that have not been fully reported before in which data from appropriate nonsaccade control stimuli are available. Subsequently, related experiments are considered in the light of these results.

6 2 1 *Measuring and predicting localization errors*

Objective and subjective foveal trajectories. In Sperling and Speelman's (1965) procedure, subjects view a display containing five marker spots (2×3 min), separated from each other by two degrees (Fig. 2a) These spots are called −2, −1, +1, +2. The observer is instructed to fixate spot −1, and then, upon an agreed upon signal (a brief dimming of the spots), to fixate spot +1. About 200–250 ms after the signal, the observer's eye will execute a saccadic movement from position −1 to the neighborhood of position +1 (Fig. 2d) Observers' 4-degree saccades are quite individualistic; some observers have considerable overshoot; and all observers show corrective saccades after about 0.2 s (Fig. 2e). The objective trajectory of the fovea as a function of time $x_o(t)$ describes an observer's objective eye movement – the physical position of the fovea as a function of time referred to an external coordinate system.

Suppose that at some time during the saccade, a flash of light occurs and that it falls directly on the fovea We ask the observer where this flash appears to fall relative to the external −2 to +2 coordinate system defined by the five spots of light. Obviously, when the flash occurs long before initiation of the movement, it is subjectively localized at location −1 (the initial fixation point); when it occurs long after the movement, it is localized at position +1 (the post-movement fixation). The subjective location assigned to a flash of light that strikes the fovea changes as a function of time; this function is called the subjective foveal trajectory $x_s(t)$, and it is measured in the same coordinates as the objective trajectory

If it happened that the objective and subjective foveal trajectories were exactly equal $[x_o(t) = x_s(t)]$ then an observer would never mislocalize a foveal flash – the observer would always report its position correctly. In general, however, the observer makes localization errors for flashes that occur during, or shortly before and after saccades. This indicates that the objective and subjective foveal trajectories are not identical. For 4-deg eye movements, the subjective trajectory is usually not quite as quick as the objective trajectory.

Procedure Some further details of the procedure are relevant. Subjects viewed the display with their heads fixed to a dental impression. Horizontal eye position was monitored by a limbus monitor (Fig. 2b,c) Initially, the monitor was dynamically calibrated with an artificial eye. During the experiment, it was calibrated between successive eye-movement trials Dynamic eye position was resolvable to an accuracy of about 3 min (for 4-degree saccades) but was recorded for subsequent analysis only to an accuracy of 12 min because, in the limbus monitor, DC accuracy is not as good as dynamic accuracy. Upon a 10-ms interruption of the display (which appeared as a dark flash), subjects were instructed to shift their fixation as quickly as possible from spot −1 to spot +1 (Fig. 2d). A thin vertical test line was flashed during the display and subjects were instructed to report its position to an accuracy of 0 1 unit of the display distance. For example, if the test line appeared to strike midway between spots 0 and 1, subjects were to report '0.5'. Preliminary experiments to teach and test the use of this method of report (during fixation) had shown that subjects could report positions to within ± 0.1 unit. The apparatus could trigger a flash when the eyes crossed the midpoint of the display or at any later time. To obtain test flashes that occur before eye movements, the apparatus is set to trigger a flash at a predetermined delay from the warning stimulus Trial-to-trial variations in this delay, together with subject variability in saccadic reaction time, yield a distribution of trials with times of occurrence before, during, and after the midpoint of the eye movement is obtained.

As a practical matter, in these experiments, all flashes occurred at precisely the same physical location, and hence at different retinal locations The subjective foveal localization was computed from nonfoveal flashes on the assumption of a rigid translation of the central 4 deg of the perceptual coordinate system (This assumption, and experiments by O'Regan (1984) testing it, are considered in detail later) As a matter of convention, time is indicated relative to the moment at which the eye crosses the midpoint of the display. Four-degree eye movements typically take less than 20 ms; therefore the movement times are within ± 10 ms of zero (Fig. 2e).

Localization errors. Some data from an eye movement experiment in which the test flash was posi-

tioned always to occur directly superimposed on the middle marker are shown in the leftmost column of Fig. 3. In this procedure, a correct report by the subject would always be '0 0' and would be indicated as 0 in the graph. This setup was chosen so that correct localization would be obvious – the test flash would appear directly superimposed on the center marker; its illusory appearance anywhere else clearly represented an error of localization. The abscissa of Fig. 3 indicates the time relative to the instant at which the eye crossed the middle marker. When the subject reports that the test flash appears displaced towards the final marker position, the error is indicated by an ordinate value greater than zero. For example, the greater-than-zero errors by subject 3 indicate that the perceptual spatial coordinates of the retina had already started to change to *their new values, even* before the eye had begun to move. Values below zero indicate that the retinal coordinates lag behind the physical movement.

Not all eye movements are precisely correct; some fall short of the intended mark and are corrected by a subsequent saccade. The data in the left column of Fig. 3 include all eye movements, those which reached their intended mark and those which did not. The advantage of mislocation errors as a dependent variable is that mislocation errors do not depend critically on the extent of the movement. It is reasonable to aggregate mislocation data from movements of somewhat different extents.

In terms of a computational theory to account for *mislocation errors*, the extent of the movement becomes a parameter, and for this purpose it is useful to restrict consideration to 4-deg eye movements which begin and end within 0.5 deg of the initial and final fixation markers. The data for each subject that describe the recorded eye positions as a function of time $x_e(t)$ can be reasonably well characterized by a 3-segment straight line: horizontal at position −1 until the start of the movement, linear slope (constant velocity) during the eye movement, and again horizontal with constant slope at the conclusion of the movement. Thereby, only two parameters are estimated, e.g., the starting time and the duration of a trajectory. These data are illustrated in the right column of Fig. 3.

The subjective foveal location (the location in space to which a test flash striking the fovea would be referred) is derived by adding the localization error $e(t)$ recorded in the experiment to the objective foveal location:

$$\hat{x}_s(t) = x_o(t) + \hat{e}(t)$$

Like objective foveal locations, subjective foveal locations $x_s(t)$ can be characterized by a 3-segment function. Fig. 3, right column, shows the observed objective and subjective eye position data for six subjects together with the 3-segment trajectories that maximize temporal prediction (minimize the

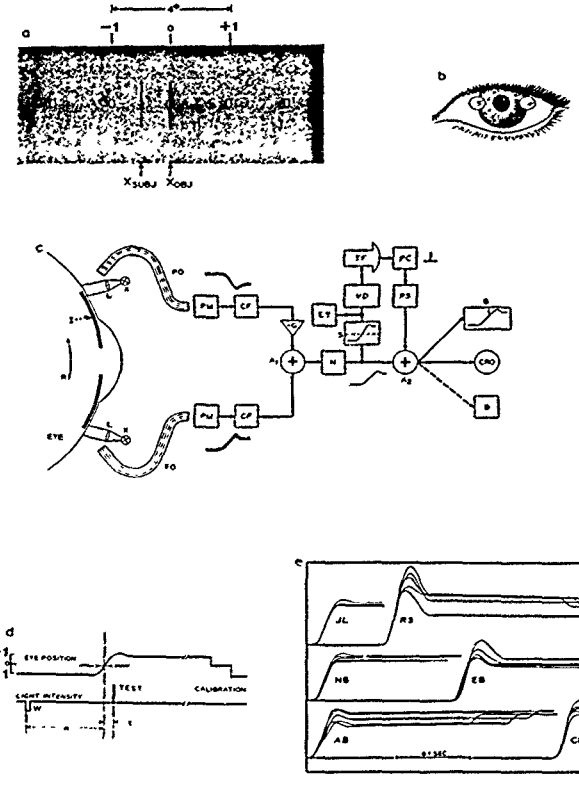


Fig. 3 (a) The display viewed by Sperling and Speelman's subjects in the saccadic movement experiments. Five bright spots of light were visible in a dark field. Initially, subjects fixed the spot labeled −1; upon a signal, they saccaded *to the spot* labeled +1. (Left/right was reversed from session to session.) Sometime before, during, or after the saccade, a thin bar centered on '0' was illuminated by a stroboscopic flash. The subject's task was to report the apparent location of the bar relative to coordinates defined by the illuminated spots. (b) A limbus monitor illuminated rectangular spots at the iris/sclera border, and measured the amount of reflected light to indicate the eye's horizontal position. (c) The limbus monitor and the linked display system. R indicates the direction of saccadic rotation, *I* is the *iris*; *X* is a *light source* focussed on the iris/sclera border by a lens L; FO is a fiber-optic bundle that collects the reflected light and carries it to a photomultiplier detector PM; CF is an electronic (cathode follower) amplifier; G is a variable gain chosen to equate the amplitude of the left and right signals; A1 adds the right and the inverted left signals to yield an estimate of horizontal eye position, *N* is a low-pass linear *filter which removes high frequency noise*. S is the threshold location in the saccade at which trigger is initiated, VD is a variable delay; TF is the stroboscopic flash gun that illuminates the target, PC and PS, respectively, are photodetectors which monitor the target and stimulus displays, A2 adds the display information to the position information to produce three records of the events of the trial, B, an ink-on-paper record, CRO, a *CRT display*; and D, a digital record on magnetic tape. (d) The sequence of events on a trial. A mon-entary darkening of the background spots is the signal to initiate the saccade; the test flash occurs at time *t*; and a calibration sequence of fixations on spots +1, 0 and −1 occurs between consecutive trials. The midpoint of the intended saccade is the point $x=0$, $y=0$, $t=0$. (e) Eye movement records obtained for six subjects. *Five traces are superimposed* in each record.

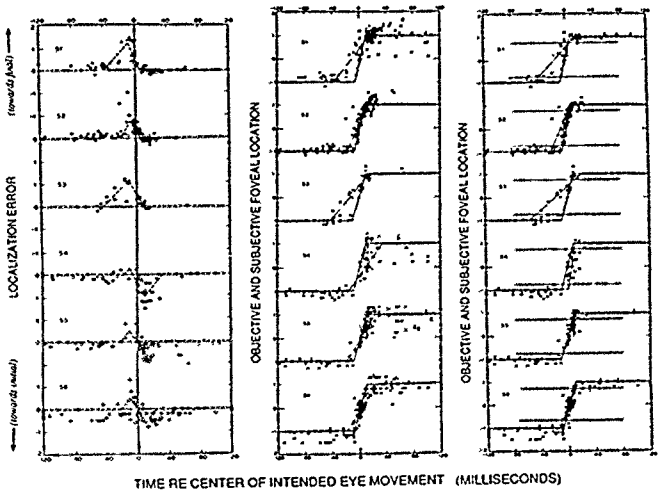


Fig 3 Left Errors in localization of a flash that occurs at location 0 as a function of its time of occurrence relative to the midpoint of the saccade The ordinate is in display units; +1 represents a mislocalization of the target at the intended final location; −1 represents a mislocalization at the initial location. Each point represents one localization judgement. The dashed lines are the difference between the best-fitting objective and subjective eye trajectories shown in center Right. Objective and subjective foveal location as a function of time relative to the midpoint of an intended 4-degree saccade Data from all 'good saccades that began and ended within 0.5 deg of the intended starting and final fixations are shown for six subjects. Each trial yields a paired objective foveal location (•) and subjective foveal location (open circle). The solid lines and dashed sloping lines minimize the squared horizontal distance from the objective and subjective location data, respectively, to the line within the middle three degrees of movement. Center. Same as right column except that all eye movements are included, and the three-segment functions minimize the vertical distance of the date to the function

root mean square (rms) horizontal distances of the data to the trajectory) in the middle 3 degrees of the 4-degree movements. This treatment of the data is appropriate for the next section, 'A purely temporal analysis of saccadic localization'.

The center column of Fig. 3 illustrates objective and subjective foveal locations for all the attempted 4-degree saccades. These data include quite a few saccades that were substantially shorter and some that were longer than four degrees The 3-segment functions which minimize the vertical rms distance of the data to the function are shown for both objective and subjective foveal locations. Strictly speaking, these objective (and subjective) foveal trajectories apply only to a homogeneous set of eye movements (e g., movements between 3.9 and 4.1 degrees) However, the differences between objective and subjective trajectories, shown as dashed lines in the left column of Fig. 3, can be used to estimate mislocation errors even for a more heterogeneous collection of eye movements because saccadic extent cancels in the process of subtracting the objective from the subjective trajectory.

The estimated durations of the subjects' eye movements (objective foveal trajectories) vary from 16 4 to 20 8 ms. The durations of the optimized subjective foveal trajectories (those which best predict localization) vary from 29 to 56 ms (mean = 45 ms). There are considerable differences in the invidividual subjective trajectories; five subjects start the subjective movement before the objective movement, and one begins later.

The difference between the straight-line trajectory estimations of the observed objective and subjective foveal trajectories is the predicted localization error. These predictions are indicated by the dashed lines in the left column of Fig. 3. The objective-subjective trajectory difference gives a reasonable account of localization errors. Section 6.3 describes a saccade simulator: the data of Fig. 3 obtained with actual eye movements will be compared to data obtained with simulated eye movements. First, however, we consider an alternative analysis of the localization judgements

6 2 2 A purely temporal analysis of saccadic localization

In the preceding section, errors in spatial localization were analysed in terms of the difference between the objective and subjective trajectories of the eye. Because localization errors are naturally measured as spatial localization errors, the spatial analysis was appropriate as an initial analysis. In this section, localization errors are considered not as spatial errors but as purely temporal errors. The reason is that spatial localization with stationary stimuli is extremely good, but, during rapid movements, small temporal errors would produce large spatial errors Therefore, it seems most likely that

the errors in spatial localization during saccades result indirectly from small temporal errors in when the saccade is estimated to have occurred relative to the test flash. Again, the temporal analysis is in terms of a mismatch between the objective and subjective trajectories of the eye but, in the purely temporal analysis, the trajectories are chosen to minimize the temporal prediction error (horizontal dimension in Fig. 3), not the spatial prediction error (vertical dimension in Fig. 3).

Temporally optimized trajectories. As was described above, ignoring overshoot, 4-deg saccadic eye movements are well described by a 3-segment function, a constant initial segment, a linear ramp, and a constant final segment. The duration of the movement, the ramp segment, is 18 ± 2 ms, and is constant for a particular subject

The subjective foveal trajectories (derived from localization judgements) also are well fitted by three-segment functions. Earlier, in the left and center columns of Fig. 3, these functions were chosen to minimize the localization error = i e , to minimize the vertical distance between the data points and the function on a graph of foveal trajectory versus time such as Fig. 3, center. Here, we are concerned with the hypothesis that all aspects of localization can be interpreted in terms of (a) temporal distortion of the movement trajectory and (b) irreducible temporal uncertainty (residual error) This requires estimating functions to maximize the goodness of temporal predictions, i e , to minimize the horizontal distance between predictions and data.

'Making' horizontal (temporal) predictions is technically more difficult than vertical (spatial) predictions Horizontal estimations can only be made in the middle sections of plots such as Fig. 3 (right). Basically, this requires selecting only good eye movements, those which finish within ± 12 min of the intended location, and making the estimates only at points between 0 25 and 0.75 of total traverse. This procedure considerably restricts the amount of data available; however, the statistics about objective and subjective location that we (Sperling and Speelman, 1965) computed for this

subset of 'good movements' were, in fact, representative of the whole.

For three of six subjects (S4–6, Fig. 3, right), the duration of temporally optimized subjective trajectories was statistically within the range of objective saccade trajectory durations (11, 20, 27 ms). For the other three subjects (S1–3), the durations of subjective saccades (33, 60, 76 ms) were incontestably longer than objective saccades. Additionally, for these subjects, the midpoint of the subjective movement preceded the midpoint of the objective movement by 6–17 ms. Because of the overall similarity of localization judgements for real and simulated saccades for all subjects, the conclusion is that different subjects perceive rapid visual motion somewhat differently, and consequently make somewhat different localization judgements. (See, for comparison, the different time courses of visual persistence measured in different subjects by Weichselgartner and Sperling, 1985).

In addition to characterizing a subjective movement trajectory in terms of its duration, there is trial-to-trial variability in localization judgements. This variability can be conceptualized as resulting either from positional uncertainty or from temporal uncertainty. Positional resolution, as measured in control experiments, was extremely good in flash localization, with errors seldom exceeding ±0.1 of the distance between markers. Therefore, we consider here to what extent localization errors can be modeled simply by temporal uncertainty in when the saccadic or simulated saccadic movement occurred.

Each objective eye position (Fig. 3, right) in the midrange between 0.25 and 0.75 of the total saccadic extent was interpreted in terms of the temporal (horizontal) deviation of the data point from the best-fitting trajectory. The root mean square deviation (rms, σ) of objective eye positions from the best-fitting objective eye trajectories varied only from 0.93 to 1.42 ms (for six subjects). Such small variations indicate that real eye movements follow a remarkably stereotypical time course. For the subjective foveal trajectories, the corresponding rms errors ranged from 4.0 to 8.2 ms, again a surprisingly small error and a small range of intersubject variation. There was no tendency for errors to vary with the duration of the subjective trajectory.

The purely temporal theory derived from the assumption that spatial localization errors in saccades are ultimately caused by temporal errors in the representation of the saccade relative to the test flash. This general principle led to the following specific conclusions concerning localization errors of a brief flash that occurred during a 4-deg saccadic eye movement with the visual reference stimuli continuously present:

(1) Flash-localization judgements in the presence of continuously present visual reference stimuli are determined primarily by visual factors.

(2) Half the subjects perceive the saccadic movement approximately correctly, and half perceive the movement to occur slightly too soon, and to be significantly longer than it actually is.

(3) Additionally, there is random temporal uncertainty with an rms value ranging from 4 to 8 ms (over subjects).

The irreducible *random* temporal errors in visual localization are somewhat greater than the temporal errors that can be estimated from Hansen and Skavenski's (1977 and 1985) motor localization tasks, which have the best experimentally observed temporal resolution. The main difference, however, is that the visual/perceptual representation of the movement is elongated relative to the objective eye movement for half the subjects in the present experiments, whereas Hansen and Skavenski (1977) found that the nonvisual (motoric) representation of the saccade in their motor task was uniformly accurate.

6.3 Simulated saccades

6.3.1 Saccadic motion smear

An eye movement simulator. To examine to what extent the visual stimulus, independent of the motor system, determined the observed localization judgements, we have to produce on the stationary eye precisely the same visual stimulus that the saccadically moving eye produces. In the case of saccades, this is a formidable technical problem which was resolved as illustrated in Fig. 4. Again, the eye movement trajectory is composed of three parts: initial fixation, movement, and final fixation, produced in three fields of a tachistoscope. The initial stimulus is shown in Fig. 4a, the final fixation in Fig. 4c.

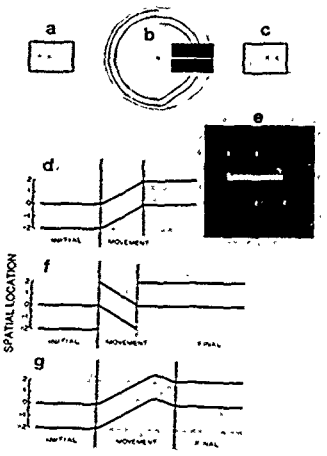


Fig. 4. Eye movement simulator: a mechanical device for generating saccadic images on a stationary eye. (a) The initial field. (b) The movement field. A large disk, rotating in the direction of the arrow, generates the moving component of the simulated saccade. The dark lines y1, y2 indicate translucent sections in the otherwise opaque disk, illuminated from behind. A masking partition M reveals only a narrow slit S to the observer. (c) The final field. (d) Representation of the position of spots y1 and y2 as a function of time for a linear saccade. The disk provides the moving field. A mirror tachistoscope, which illuminates only one field at a time, coordinates illumination of the movement field with the stationary initial and final fields. (e) The motion smear: a time-exposure photograph of the stimulus generated by the movement field. Also shown, markers in their initial and final positions, and the test flash. (f) A reversed eye movement. The direction of the motion segment is reversed relative to the initial and final fields. (g) A simulated saccadic trajectory with overshoot.

The moving stimulus was produced by a rotating disk as shown in Fig. 4b. On an otherwise opaque disk, two transparent, approximately concentric, curves $c_1(\rho, \theta)$ and $c_2(\rho, \theta)$ are drawn. The disk spins continuously, is illuminated from behind, and is viewed through a narrow slit arranged along a horizontal radius. The observer sees two spots of light which move left or right as the radial distance ρ varies with θ. Almost any movement trajectory of the spots can be produced by appropriate choice of $c(\rho, \theta)$. When the disk spins at 6.3 rotations per second, the trajectory occupies 20 ms. Therefore, illumination of the moving field (Fig. 4b) has to be coordinated by elaborate temporal synchronization and optical superposition with illumination of the stationary initial and final fields (Fig. 4a and c).

A trial begins with the initial field displayed continuously. When the subject presses a key, at the first available rotation after 0.5 s, the initial field shuts off and the moving display turns on for 20 ms. After the movement section has passed, its illumination is shut off and the final field 3 is illuminated. In this setup, it was feasible to present only 2 of the 5 marker spots of the previous experiment; the −1 and +1 markers, corresponding to the initial and final fixations. A time-exposure photograph of the moving portion of the display is shown in Fig. 4e, as well as the initial and final fields. It illustrates the quite uniform motion smear produced by the simulated eye movement.

Finally, a Risley prism was inserted into the view path. By rotating the prism, the left/right orientation of the entire display could be reversed. By reversing the static initial and final positions and also reversing the entire display orientation (Fig. 4c), only the direction of movement remained reversed. Four conditions of movement trajectory

were investigated (a) normal motion trajectory (Fig. 4d), (b) reversed motion trajectory (Fig. 4f) (c) sampled motion trajectory – illumination of the moving field is turned off during the interval between initial and final fields – and (d) double reversed-direction sampled motion trajectory, i.e., this looks the same to the observer as the normal sampled motion trajectory. The temporal sequence of the normal initial and final fields is reversed, and the display is then reversed spatially with a Risley prism. This is a check for any undetected difference between initial and final fields that might affect the reversed motion trajectory. This apparatus was used to study continuous and sampled movements of various lengths and durations.

6 3.2 The appearance of motion smear

While the main purpose of the apparatus was to study spatial localization I digress for a moment to consider the subjective appearance of the saccadically moving stimuli. The most surprising observation was that, even though the stimulus consisted of bright spots viewed against darkness, observers did not spontaneously discriminate the correct from the reversed movement trajectory when the imposed movements occupied 20 ms and the total distance traversed ranged from one to four degrees, according to the viewing distance. With the eye movement simulator, visual sensitivity to the various aspects of the retinal movement trajectory could be isolated. On alternate presentations, the illumination of the movement field was turned off, thereby eliminating the movement smear entirely and substituting 20 ms of darkness. Naive viewers did not spontaneously notice any difference between consecutive displays of continuous motion and sampled movement. Indeed, observers did not report a difference in the appearance of motion smear even when they were pressed, although many irrelevant aspects of the displays caught their attention.

When the difference between the normal-, reversed- and no-smear (sampled) displays is pointed out to viewers, they can notice a barely discriminable difference in the motion smear between continuous-movement and the sampled-movement trials but the direction-of-movement discrimination appears to be impossible. Of course, when the speed of the motion trajectory is slowed down by a factor of 10 or so, all the appropriate relations can easily be perceived.

Visual masking of motion smear. When the motion smear occurs alone (i.e., only Fig. 4b is shown and the initial and final fields (Fig. 4a and c) are turned off), the smear itself (Fig. 4e) is quite easy to detect. But the discrimination of motion direction in the smear remains difficult. MacKay (1970a), Campbell and Wurtz (1978) and Corfield et al. (1978) propose that the difficulty of detecting motion smear is due to visual masking by the visual stimulation that immediately precedes and follows the smear. In one experiment (Campbell and Wurtz, 1978), subjects initiated eye movements in the dark. During the eye movement, a light was turned on. When the light remained on only very briefly, subjects reported that the scene illuminated by the light was clearly visible and sharp (thereby reproducing once again the observation of clear vision during saccades). As the light remained on for 20 ms and longer during a long saccade, the scene appeared to become extremely blurred, much like the motion smear represented in Fig. 4e. On the other hand, if the light remained on for more than 40 ms *after* the saccade ended, the previous saccadic motion smear became invisible. Thus 40 ms of post-saccadic stimulation masked saccadic motion smear.

To study motion smear in the stationary eye, Corfield et al. (1978) represented saccadic motion smear by a stationary blank field, like that of Fig. 4e. They preceded and followed the blank field by different textured fields, principally sinusoid gratings and combinations of sinusoids. When the blank field was as brief as it would have been during normal saccadic eye movements, the blank field was not visible – it was completely masked by the preceding and following stimuli. While various parameters of masking were investigated, the process of visual masking itself was not elucidated in these experiments. However, whatever the masking process in the stationary eye may be, the experiments demonstrate that it is also sufficient to account for the invisibility of motion smear in the saccadically moving eye.

6.3 3. The effect of motion smear on spatial localization

Spatial localization during imposed saccade-like movements. The difficulty of observing motion smear during real and simulated saccades even under optimal conditions for its appearance (i.e., a scene consisting of bright spots on a dark background) suggests that motion smear would not contribute to other kinds of psychophysical judgements. Nevertheless, I (Sperling, 1966) used the saccadic motion simulator shown in Fig. 4 to investigate spatial localization. The motion trajectory on the retina and the psychophysical procedure were analogous to those of the saccadic spatial localization task described above in section 6.2.

The saccadic movement simulator was arranged to provide the linear saccadic movement stimulus illustrated in Fig. 4d. During the movement trajectory, a thin test line was flashed briefly, and the subject was asked to localize the flash *relative to* the position of moving spots at the instant of the flash. The localization judgement was similar to that illustrated in Fig. 2a, with the subject initially fixated on the spot labeled −1, except that the only other spot visibe was +1. The subject maintained fixation rather than moving his eyes, and the display moved quickly so that, after the movement, the spot +1 was at the fixation point.

The movement trajectory was a linear translation between initial and final positions which traversed the distance in 20 ms (Fig. 2d). In this experiment, the viewing distance was increased so that the total length of the movement trajectory was 95 min of visual angle. The movement was somewhat slower than a natural saccade and somewhat shorter than the saccades studied in the previous section. These experimental parameters were chosen to maximize the number of localization judgements

between – rather than at – the end points in order to give the imposed trajectories the maximum opportunity to exert differential effects. Normal continuous movements, reversed movements (Fig. 4f) and sampled (no smear) movements were tested. The test line, rather than flashing at different retinal locations, as in the saccadic movement experiment, always occurred displaced slightly from the fixation point (the fovea) in the direction of the simulated movement (0.2 of the movement distance as shown in Fig. 4e).

The results can be described succinctly. No subject showed any significant difference between the normal and reversed movement conditions with respect to spatial localization of the test flash. In other words, the direction of the saccade-like image smear did not influence spatial localization judgements – they were determined by the pre- and post-movement fields.

There were minor localization differences between continuous-movement (normal or reversed) and sampled movement for some subjects and not for others. When there was a difference, it was a greater tendency in the sampled movement displays for the test flash to be localized at the initial or final positions, not in between. Data representing each type of performance are shown in Fig. 5. For example, S4 hardly ever localizes a flash at locations between −1 and 0 in the sampled control (top and bottom panels, Fig. 5), but does so frequently in both smear conditions (middle panels, Fig. 5).

Spatial localization during much slower-than-saccadic simulated movements. The saccadic simulator can be used to impose retinal motion at velocities that are higher or lower than saccadic velocities. Subjects viewed the moving stimuli while maintaining fixation on a stationary fixation point. In order that subjects did not correctly anticipate the image motion and move their eyes, the direction of motion was random from trial to trial. As in the case of movements at saccadic speeds, subjects were asked to judge the location of a brief flash relative to the moving coordinate system defined by the dots.

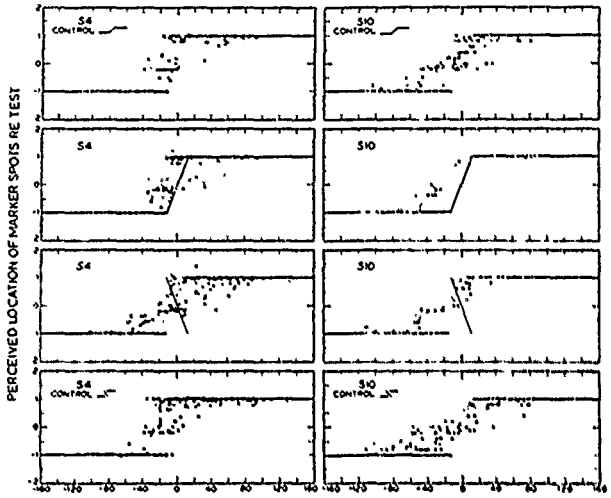


Fig 5 Localization data from an experiment comparing simulated normal, reversed and sampled eye movements. The abscissa is time relative to the midpoint of the movement trajectory; the ordinate is the perceived location of the marker spots as computed from the judged relative location of a brief test flash. Solid lines indicate the actual location of the marker spots. Data from two subjects are shown, each point represents a single judgement. In the sampled movement conditions ('controls'), all illumination was turned off during the 'movement' section of the display. In the top control condition, the initial and final marker spots were those used for the sampled motion condition, in the bottom control condition, the initial and final marker spots were those used for the reversed motion condition, the actual direction of motion was the same in all conditions (see text).

The left half of Fig. 6 shows the localization of the test flash relative to the moving coordinate frame during 4-deg continuous movements, with durations of 50, 125, 250 and 500 ms. The right half of Fig. 6 shows localization during no-smear (sampled) movements. As in the previous experiment, the test line occurred slightly displaced from fixation in the direction of movement (Fig. 4e).

As the duration of the movement component of a trajectory is stretched out in time so that it is much slower than saccadic eye movements, there eventually come to be quite obvious differences in appearance between a display with a normal movement trajectory and one with the movement trajectory turned off. The trajectory between the initial and final position is correctly perceived as a slow translation; when it is absent, a blank period is perceived (e.g., blank times of 250 and 500 ms are quite obvious). With long blank times, the localization task for sampled-motion (interrupted) stimuli is ambiguous. When the localization task is redefined as 'judge the flash relative to where you believe the markers would be if they were visible,' the results of localization experiments with continuous-movement and with sampled-movement stimuli are quite similar. (Localization judgements made when the coordinate frame appeared to be invisible at the instant the test flash occurred are indicated by x in Fig. 6.) Because the linear interpolation of visual motion is so natural, even when there is an obvious blank period between initial and final positions, the localization task is not appropriate for revealing the differences in appearance.

Localization judgements during movements of various speeds yield some obvious and predictable results and some surprising ones. (1) For practical purposes, flash localization judgements do not distinguish between continuous and sampled motion

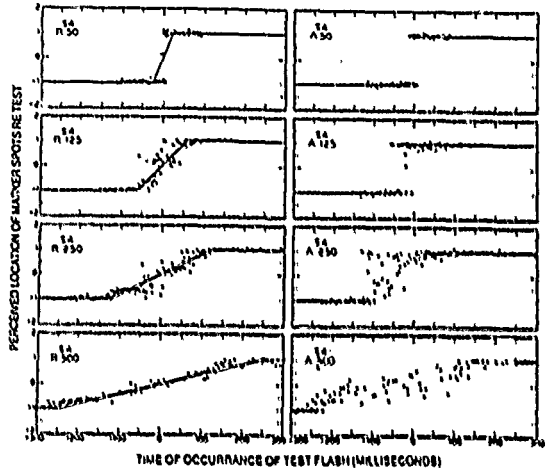


Fig 6 Spatial localization compared in continuous and sampled motion stimuli of various speeds. The solid lines indicate the actual motion trajectory; the coordinates are as in Fig. 5. Left panels show continuous motion stimuli, right panels show sampled (sometimes called apparent) motion stimuli. Open circles indicate the perceived location of the moving marker spots as computed from the judged location of a brief test flash; x's indicate that the marker spots were temporarily eclipsed at the instant of the test flash, and the location judgement was made relative to an inferred trajectory.

at saccadic and near-saccadic velocities, i.e., 1.6 deg, 20 ms imposed movement (Fig. 5) and the 4.0 deg, 50 ms imposed movement (Fig. 6). (2) When markers spaced by 4 degrees traverse the intermarker space in 50 ms, flashes are nearly always localized only at their initial and final locations. This strictly bimodal distribution of localization judgements obtains for greater spaces or quicker traverses. (3) For motions of 125 ms (less than 1/6 of saccadic velocity), the localization judgements for continuous and sampled movements are profoundly different. For continuous movements, test flashes are localized quite accurately along the continuous motion trajectory. For sampled movements, test flashes are localized only at the endpoints of the trajectory. Bimodal localization judgements accurately reflect the fact that the motion trajectory is bimodal. However, bimodal localization judgements fail to reflect the phenomenology: the 125-ms sampled motion stimulus appears to jump across the space and to take a non-zero interval of time doing so. (3) With long blank times, localization judgements made during the interval in which the moving reference stimulus was turned off indicate remarkably accurate predictions of position. While these 'cognitive' localization judgements based on an invisible stimulus are somewhat more variable than 'perceptual' judgements based on an actual moving stimulus, they are nevertheless remarkably similar.

Spatial localization with overshoot trajectories. Saccadic trajectories typically have brief overshoots at the end of the movement in which the line of sight briefly extends beyond the intended endpoint and then returns back to the steady intersaccadic position. One of our subjects typically showed overshoots as large as 50 percent of the movement. To determine whether such a trajectory might influence perception, we (Sperling and Speelman, 1965) simulated a trajectory with 33% overshoot on the movement apparatus (Fig. 4g). In order to give the unusual trajectory the maximal opportunity to affect localization, it was run much slower than real time: the linear component from start to first arrival at the final location was 80 ms, and the overshoot duration was an additional 40 ms (see Fig. 4g). Stimulus conditions were similar to the previous simulated-motion experiments. As before, the subjects' task was to judge the location, relative to the moving coordinate frame, of a test flash that occurred during the movement.

Two motion extents (from initial to final) were investigated: (1) 95 arc min, and (2) 9.4 arc min. This small extent of movement was produced by means of an inverting telescope which reduced the 95 arc min by a factor of 10. Data for one typical subject are shown in Fig. 7. Two aspects of the data are noteworthy. Test flashes are never localized outside the initial-to-final interval in spite of the large, simulated saccadic overshoot. Indeed, even with these slow simulated movements (which are at least 3 times slower than real time) there is no evidence that the overshoot has any effect on perception.

The stimulus to study overshoots also revealed another important characteristic of localization. With the small displays (and correspondingly slow, small movements), there was no difference between localization in continuous-movement displays and in sampled movement displays with the motion segment turned off. However, with the 1.6-deg movement, the difference between continuous and sampled motion is overwhelming: test flashes during continuous movements nearly always appear at locations between the initial and final markers; with sampled movements, the test flashes are localized only at the endpoints. These data extend the top panels of Fig. 6, showing that the continuous-sampled motion difference can be made to vanish with sufficiently small displays, just as Fig. 4 showed that the continuous-sampled motion difference vanished with sufficiently brief (20 ms) motion periods (Fig. 4). All these differences between continuous and sampled motion follow immediately from a Fourier frequency analysis of the stimuli (Watson et al., 1986). Obviously, the parameters of visual motion exert a controlling influence on how test flashes are localized relative to a saccadic-like moving reference stimulus. This will be important in interpreting test flash localization during real saccades.

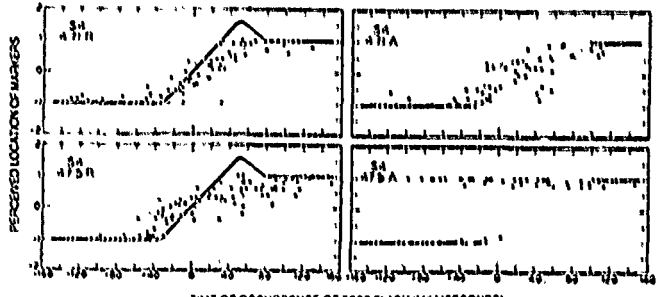


Fig. 7. Location judgements with continuous and sampled motions of two sizes. In the lower panels, the moving markers traversed from 47.5 min on one side of fixation to 47.5 min on the other side. In the upper panels, the display size was reduced by a factor of 10. The movement was a much-slower-than-real-time simulated saccade with overshoot (see text). Overshoot appears not to matter at these motion speeds. Turning off the movement field (saccadic *smear*) matters only for the larger movement.

All these observations indicate that the invisibility of high-velocity motion smear, which is present naturally during saccades, and errors of test flash localization relative to continuously moving images are not unique to saccades, but reflect visual responses to moving retinal images. With respect to motion smear, the visual system may have developed insensitivity to rapid-motion smear as a way of dealing with saccadically induced stimulation. However it may have evolved and developed, the visual system that we now have seems to respond in the same way to the same retinal stimulation – whether the stimulation is saccade-produced or object-produced. The following sections investigate the extent to which the mechanisms of spatial localization, which include nonvisual mechanisms, operate similarly in real and simulated saccades.

6.3.4. *What determines the subjective foveal trajectory x(t) during eye movements?*

Comparisons of spatial localization in real and simulated saccades. The saccadic simulator was designed to imitate the retinal stimulation produced by objective movements. While it is possible to reproduce the trajectory of any particular eye movement, it is not practical nor necessary to reproduce the normal variation of saccadic trajectories in this experiment. Since we now know that saccadic motion smear does not importantly affect visual localization judgements, it is sufficient to use one typical trajectory in the imposed-motion control experiment – a 20 ms, 4 deg, linear motion trajectory. How does spatial localization with this simulated eye-movement trajectory on the stationary retina compare to localization during real saccades?

To compare real and simulated saccades, a subset of all the saccades was selected for which the simulated trajectory was a good approximation, i.e., saccades that began and ended within 12 min of their intended locations. For each saccade in this subset of 'good' saccades, the retinal location of the test flash relative to the saccade was noted. In the simulated saccade condition, the test flash was produced at precisely the same retinal location relative to the

moving reference points, and the observer made the corresponding location judgement of the test flash relative to the marker spots.

In the simulated saccade experiment, the subject sees two stationary spots (−1, +1, Fig. 2a) with a 2-degree separation between adjacent spots. When the subject is ready, the subject initiates a trial by pressing a button. After a variable delay, the lines quickly move to replicate the saccade-induced sweep of the grid across the retina. During this sequence of events, the brief test flash occurs, and the subject's task is to report the location of the flash relative to the moving coordinate system.

Twin procedure. Each trial in the simulated saccade experiment is the twin of an earlier saccadic trial. The stroboscopic flash in the simulated movement trial always occurs at the same coordinate point relative to the moving reference stimulus (0,0) and the same time relative to the movement as did the original flash in the saccadic trial. The order of the simulated trials duplicates the order of the original 'good' saccadic trials. The perceived location of the brief flash relative to the saccadically simulated moving coordinate system on the stationary eye is precisely analogous to the flash location relative to the real saccade-induced moving coordinate system. That is, if localization with eye movement and simulated movement trajectories were entirely equivalent, observers would make precisely the same localization judgements in the two conditions.

Five of the six subjects whose saccadic localization data were shown in Fig. 3 served in the simulated saccade experiment. On the whole, the localization data from the simulated eye movement condition were similar to the data from real saccades. The main significant differences were a few instances in which observers judged a test flash in the simulated display at an endpoint when they had previously, in the real saccade, judged the flash at an intermediate position. Were these differences due to residual physical differences between the real- and simulated-movement images? For example, the display for the real saccadic movements (Fig. 2) showed five marker spots, whereas the simulated saccadic movement image showed only the −1 and +1 marker spots (Fig. 5).

Whether the slight differences in localization judgements between these real and simulated movements were due to residual differences between the displays in saccadic and simulated conditions or to differences in the process of localization between real saccades and simulated saccades was not determined for 4-degree saccades. However, if there are differences in localization between real and simulated 4-deg saccades, they are certainly not much larger than the trial-to-trial and subject-to-subject variability for the displays studied here.

To help resolve the issue of saccadic versus simulated saccadic localization differences, data were obtained with one subject using 8 (rather than 4) degree real and simulated saccades. With the larger motion extent, the localization differences between real and simulated saccades increased strikingly; there was much more localization of the test *between* endpoints for the 8-deg real saccade, and more localization *at* the endpoint for the simulated saccade. The change in simulated saccadic localization with scale followed basically the pattern illustrated in the sampled motion conditions of Fig. 7. Another indicator of a change in the localization process with saccadic size was that the duration of the 8-deg subjective trajectory $x_{ss}(t)$ was briefer than the duration of the 4-deg subjective trajectory $x_{ss}(t)$, although the objective 8-deg saccade took about twice as long as the 4-deg saccade. While such data were obtained with only one subject, they suggest that, even when visual information is prominent in the visual field, in larger saccades the motor movement itself importantly influences localization judgements in larger saccades. This issue is considered in the next section.

Visual versus nonvisual factors. As noted above, dynamic visual stimulation during saccades – certainly during brief saccades – is unimportant because it is effectively masked by stimulation arising from the static pre- and post-saccadic fields. Therefore, the static pre- and post-saccadic stimuli are the primary contributors to dynamic visual localization judgements, a matter illustrated in Fig. 5 (insensitivity to trajectory) and whose consequences will be taken up in more detail later. By contrast, nonvisual (motor system) factors might become more prominent in localization for large eye movements (e.g., 8 deg for the kinds of displays considered above; see also Pola, 1972). In addition to size of a saccade, the most obvious factor that determines the role of nonvisual factors is the visibility of the visual scene itself. As the scene becomes dimmer and thereby less visible, and ultimately invisible, spatial localization of a test flash will be determined more and more, and finally exclusively, by nonvisual knowledge of eye position.

A number of experimenters, among them Bischof and Kramer (1968), Mateeff (1978) and MacKay (1970b, 1973), have studied spatial localization during saccades in the presence of background stimuli. O'Regan (1984) points out, in effect, that the failure to use an adequate simulated saccadic control may invalidate conclusions about possible localization mechanisms from these earlier studies. In his own experiments, O'Regan (1984) did use a simulated saccadic control. While he found similarities between real and simulated saccades, his data were not collected in a twin procedure and they do not allow one to detect small differences that might exist. Even so, O'Regan's data, like the data reported in the previous section, show a tendency to localize test flashes at intermediate points more often for real than for simulated saccades.

The conclusion, based on the data described here and from published data in which judgements were made of the spatial location of a visual test flash against a structured visual background, is that, in normal viewing, visual factors predominantly determine test flash localization, whether the eye or only the image is moving. However, when visibility is reduced (darkness is an extreme case) or when nonvisual components are enhanced (as in large eye movements), or when the test flash is judged relative to the body rather than relative to another visual stimulus (Hansen and Skavenski, 1977, 1985) then nonvisual factors become more and more important.

6.3.3. *Visual localization in the dark*

Spatial localization after a visual frame of reference is extinguished. What is the perceived location of a test line flashed while the eye is executing a saccadic movement in the dark? For example, the observer views a fixation point *A*. It goes off and subsequently a point *B* appears which is the target of an intended saccade. The observer is instructed to saccade from *A* to *B*, but the *B* is turned off 100 ms after it appears, before the eye begins its movement. A stroboscopic flash is programmed to strike the fovea sometime before, during, or after the movement. The observer perceives this flash as originating from some point in the environment. Where? And how does the observer indicate where?

The test flash is judged relative to the memory of the extinguished markers, *A*, *B*. Obviously, just as in the light, a flash striking the fovea would be localized first at the initial location and then, sometime after the end of the saccade, at the final location. During, and shortly before and after the saccade, flashes are localized at intermediate points. For this experiment, there is no equivalent control experiment in the stationary eye. Spatial localization in the absence of visual stimuli must be controlled by nonvisual factors. The computation of the location of the fovea could rely on efferent outflow signals or upon proprioceptive feedback, but, in the absence of visual stimuli, it cannot rely on vision. This procedure allows one to measure the quality of visual localization information that is available via the motor system.

Experiments show that the perceived location of a foveal flash during a saccadic movement in the dark changes very slowly relative to the speed of perceived location of a foveal flash viewed against a visually structured environment. It takes hundreds of milliseconds for the perceived location of flashes in the dark to move from *A* to *B* compared to the few tens of milliseconds needed to execute the saccade (see Matin, 1986, for a review).

Hypothesis: feedback determines the frame of reference In a closely related experiment in which the task of the observer is not to judge the test flash relative to a previously remembered location but to strike the flash directly with a hammer, Hansen and Skavenski (1977, 1985) find extraordinarily accurate localization. In their experiments, subjects have almost perfect, almost instantaneous, nonvisual knowledge of position relative to the body of the saccadically moving eye. Why does this knowledge not manifest itself when the subject is asked locate a visual flash relative to a remembered visual location? Skavenski (Ch. 5) proposes that the subject adopts a different frame of reference in the two tasks.

Adopting a task-dependent frame of reference means that the weight the subject assigns to different sources of information in arriving at a response depends on the task. This matter will be considered in the next section. Here we note that for the subject to achieve an optimal weighting of information sources in any of these complex tasks requires practice with feedback. All humans practice all their lives coordinating saccadic eye movements with body movements. Therefore, it is not surprising to discover that a subject is aware of the position of the eye relative to the body. However, people never practice making a purely visual judgement, which does not involve any body movement, in circumstances in which illumination is suddenly extinguished. The hypothesis put forward here is that, because all information necessary to perform this task is available to the subject, with practice a subject should be able to learn the purely visual task. The critical aspect for all the visual localization experiments considered here is that they measure what the subject habitually does, not the subject's capacity. To make an inference about capacity – an intrinsic *inability* to visually localize stimuli when a visual frame of reference is extinguished – requires that a subject *fails to learn* in an experiment with feedback (Sperling et al., 1990).

6.4. Models for spatial localization during eye movements

This section considers theories for the data that have been presented concerning the localization of flashes which occur during or proximal to eye movements. (The issue of how information from successive saccades is combined is considered in a later section.) Basically two kinds of information are involved: purely visual and visuo-motor. Visual information ('retinal factors' in the literature) refers exclusively to information carried by light, and it is the same in real and in simulated eye movements. Visuo-motor information includes efferent or afferent motor-system information ('extra-retinal factors') linked to vision. Here *visuo-motor* information is abbreviated to *nonvisual* although it makes no sense to consider nonvisual information alone – without vision – in a visual localization task.

Obviously, only nonvisual factors distinguish between real and simulated eye movements. (This requires that the simulated events are truly equivalent to the retinal images during eye movements, a technical requirement that has often been violated because it is difficult to achieve.) To evaluate the role of purely visual and visuo-motor factors in spatial localization, it is useful to have in mind at least one specific model for how each kind of information might be processed. We consider here a model for each process.

6.4.1. Model for purely visual localization during saccadic-like image sequences

Attention gating model: events, glimpses, episodes The mechanism of the attention gating model of Reeves and Sperling (1986) and Sperling and Weichselgartner (1989) is the core building block of all the proposed models. The gating model deals with the mental representation of external stimulus events that occur in close temporal and spatial proximity to each other. According to the model, such closely contiguous events are not stored or accessed individually in memory; they are packed into an attentional *glimpse*. The glimpse is the smallest attentional unit. Its contents result from a single opening and closing of an attentional gate. The events that comprise a glimpse define its space-time window, much as the location where a camera is pointed and the time its shutter is open determine the space-time window of a photograph. A glimpse may incorporate events that span from about 1/4 s up to about 1 s. One or more glimpses may be bundled into an episode. The episode is the unit that is accessed when information is retrieved from long-term memory.

The gating model (Reeves and Sperling, 1986) describes the computational mechanism that creates attentional glimpses. In the environment in which it was developed and tested (highly controlled stimulus sequences which give the experimenter full control over the sequence of events) the gating model has great predictive power. The time course of attentional glimpses was found to be highly constrained and constant for an individual. The basic premise of the attention gate is that the *amount* of information recorded internally about an external stimulus event is proportional to the amount of attention received by the stimulus event at its time of occurrence, which, in turn, is determined by where within the glimpse the event occurs. The amount of information recorded about an event is characterized by a positive real number, its *strength*. The computational concept of *strength* represents the structural concept of the *strength of a link* which connects an event to the node designating the episode to which the event is attached.

To arrive at an observable response, information recorded in a glimpse must be interpreted. Here, the following assumptions are made. In the decision algorithm, information is weighted according to its strength. In arriving at a decision, it is not critical whether information is acquired from one or from several glimpses.

According to Weichselgartner and Sperling (1987), there are two kinds of glimpse, automatic and voluntary. In automatic glimpses the opening of the attentional gate is determined by the contents of the glimpse itself. For example, in the glimpse that records the test flash in a localization experiment, the attentional gate admits information about the test flash itself, about the visual field in proximity to the test flash, and about other events that may have occurred in close temporal proximity to the test flash.

In a voluntary attentional glimpse, information is admitted according to an attentional gate which is triggered by events external to the glimpse itself. For example, if the occurrence of a test flash were a cue to the observer to begin remembering the configuration of a complex background, the test flash would be remembered in the first glimpse and background information would be recorded in a subsequent glimpse.

Attentional gating functions are derived empirically from experiments in which a subject is asked to remember a single test event embedded in a visual field populated with other events (Reeves and Sperling, 1986; Weichselgartner and Sperling, 1987). The spatial and temporal location of events that the subject extracts along with the test can be used to derive the spatio-temporal gating function. Fig. 8 illustrates the time courses of an automatic and voluntary glimpse derived from such an experiment. The trigger event was the brief flash of an outline square, much like the test flash in the spatial localization experiments. The other events that the observer attempted to remember along with the outline square were flashes of letters, superimposed in a rapid stream inside the square. The voluntary attentional glimpse was quite slow compared to an automatic attentional glimpse – just as slow as if it had required a spatial shift of attention. Such voluntary attentional glimpses have been studied in a great variety of contexts (Reeves, 1977; Sperling and Reeves, 1980; Reeves and Sperling, 1986) and been found to be remarkably constant for a particular individual over a variety of conditions. The properties of the automatic attentional glimpses have not yet been quite as well defined.

Sources of information for localizing test flashes during saccade-like image movements. In a localization experiment, there are four stimulus events: (1)

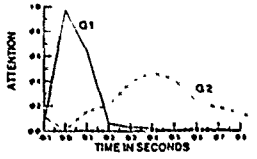


Fig. 8. Time course of attention triggered by the occurrence of a brief flash of an outline square. Two glimpses are distinguishable: G1, an automatic glimpse which records the flash and closely neighboring events, and G2, a controlled glimpse which mainly records events 200–600 ms later. The abscissa is time relative to the trigger event which occurred at time 0.0; the ordinate is the instantaneous amount of attention. (After Weichselgartner, 1984, Fig. 4.19a.)

the premovement field (a visual image of the retinal stimulus as it existed before the movement); (ii) the movement flow field (a representation of the movement and, possibly, a visual respresentation of motion smear caused by the retinal sweep of the background), (iii) the post-movement field (an image of the retinal stimulus after the movement), and (iv) the test flash to be localized. When the test flash occurs during or in close temporal proximity to the movement, these events are represented in three glimpses: (1) a controlled glimpse which contains the premovement field plus weak representations of subsequent events, (2) an automatic glimpse triggered by the test flash which contains the test flash and possibly all the other events, pre-field, movement, and post-field, and (3) a controlled glimpse which contains primarily the post-field. It is assumed that, at the level of processing where a localization judgement is made, the automatic episode in which the test flash is recorded will be given the primary weight in determining the localization judgment.

Content analysis of the visual component events There are two separate kinds of motion information: motion smear (a static retinal image) and the *motion flowfield*. In the case of saccadic movements, because of masking by pre- and post-saccadic fixation fields, motion smear is usually invisible and therefore unavailable for any subsequent processing (see section 6.3.2.). Therefore, motion smear is retained explicitly as 'null' in Glimpse II. In the case of test flash localization in much-slower-than-saccadic movement, the motion smear takes on non-zero values.

The motion perception system calculates the motion flowfield, an assignment of motion velocity (a vector) to each point in two-dimensional space (Clocksin, 1980, Hoffman, 1982; Koenderink and van Doorn, 1986; Longuet-Higgins and Prazdny, 1980, Sperling et al., and Perkins, 1989). The neural computations of image motion are assumed to be the same whether the flowfield is produced by saccades or by image movements and whether the motion is continuous or sampled (e.g., Adelson and Bergen, 1985; Heeger, 1987; van Santen and Sperling, 1984, 1985; Watson and Ahumada, 1983; Watson et al., 1986). In the automatic glimpse containing the test flash, the temporal and spatial proximity of the saccadic image-velocity vectors to the test flash determine the weight of image motion in the localization judgment. This weight, expressed by the area (w_3 in Fig. 9a) under the test flash's attention glimpse function, represents the belief: 'The test flash occurred during the movement'. The computation of the weights of the various events that comprise the test-flash glimpse is illustrated in Fig. 9.

Visual persistence and test flash localization. While pre- and post-saccadic fields mask saccadic motion smear, the test flash in localization experiments is typically not masked. Therefore, because of its visual persistence (e.g., Efron 1970a,b; Sperling, 1960, 1967), the test flash becomes a de facto component of subsequent stimuli. Unless the test occurs long before the movement, it will persist into the post-saccadic background stimulus. In computing spatial relationships within an image, the persisting visual image of the test flash is treated in the same way as a physically present test image would be. Therefore, the persistence of a test flash

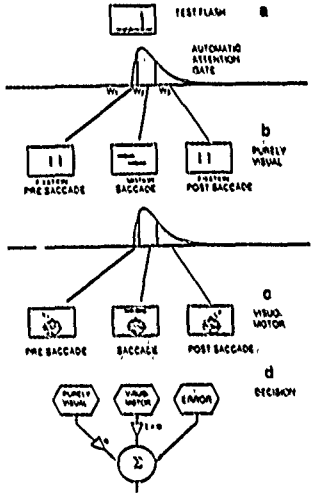


Fig. 9. Models for purely visual localization, visuo-motor localization, and for the combination of different sensory cues. (a) The intensity of a nearly instantaneous test flash as a function of time, and the attentional window of the episode in which the flash is recorded. (b) Representation of the retinal images of a stationary background before, during, and after a saccade. In purely visual localization, the images, weighted by their proximity to the test flash (w_1, w_2, w_3), are the only information available for the judgement/decision process. The particular weights shown here are appropriate for a test flash that occurs early in a saccade. (c) In visuo-motor localization, the representations of the pre-saccadic eye position, the moving eye positions and the post-saccadic eye position are weighted to derive the eye position to be linked with the test flash. (d) A simple model for the combination of cues from different sensory sources that has been found to accurately predict the data from multi-cue experiments. Data from various sources, weighted according to the task requirements ($0 \leq a \leq 1$) are summed together with irreducible sensory noise (*ERROR*) to generate a response on each trial.

at particular location of the post-saccadic field is interpreted as very strong evidence (large weight) for occurrence of the test at that location. Visual persistence is probably the most important factor in the tendency of most observers to locate test flashes that occur shortly before saccades relative to the post-(rather than pre-)saccadic visual environment (O'Regan, 1984).

The entire rise and decay of visual persistence was mapped for individual subjects by Weichselgartner and Sperling (1985). Large individual variations in visual persistence functions were found. It would be of great interest to determine whether the duration of an individual's persistence function is correlated with the tendency toward localization errors in the post-saccadic direction.

When multiple tests are flashed during a saccade, their location is judged as though they all occurred at the same time (Matcoff, 1978; O'Regan, 1984).* That is, they are judged at their actual retinal positions without taking into account that the eye and the background retinal image were moving. This is explained by the persistence of all the flashes in the post-saccadic image. The simultaneous persistence of many flashes allows their retinal spatial relationships to be computed in the same way as if they had actually occurred at the same time. Such internal evidence of consistent spatial relationships is compelling, and overwhelms evidence of a differential time of occurrence during movement.

Nondistortion of images flashed during saccades. It was mentioned above that complex images briefly flashed during eye movements do not appear distorted. This observation has a long history (e.g., Woodworth, 1906, 1938; Campbell and Wurtz, 1978). One might expect images to appear distorted because the latency of visual responses varies with retinal location, as well as with stimulus contrast and with spatial frequency. In the stationary eye, this variation in the precise arrival time of sensory information at some more central process-

* Matcoff has a different interpretation of his data.

ing center might be irrelevant. In the saccadically moving eye, small time differences represent large location differences. Again, it is not that time differences in sensory processing are somehow compensated for during saccades. When stimuli are flashed in isolation, retinal location and intensity influence perceived location (O'Regan, 1984). However, in the face of visual persistence that is long relative to the difference in onsets, there is a long period when all the stimulus information is simultaneously available. As with test flashes produced in rapid succession during saccades, the computation of spatial relationships between simultaneously available stimulus components provides evidence which is given overwhelming weight relative to evidence of difference in arrival times. Small differences in arrival times are used with extraordinary sensitivity in motion computations. During saccades, however, object motion is *masked* by saccadically produced image motion and perhaps also *suppressed* (see below, section 6.5). Small arrival time differences *un*associated with stimulus motion are not generally treated as significant data. It requires unusual stimulus conditions to demonstrate that uncompensated sensory processing times can enter into perceptual computations.

Saccadic versus simulated saccades In the preceding discussion, it has been taken for granted that statements made about moving images in saccades would hold for simulated saccades in the stationary eye. For example, flashing a complex scene during a saccade from one marker spot to another is simulated by flashing a complex scene during the movement of two marker spots on a stationary retina. It seems so obvious that moving two marker spots would not alter the appearance of the flashed scene that it has not been explicitly tested. Similarly, it seems likely that two quickly consecutive flashes striking the fovea during the sweep of a moving background would both be localized at the same place. But this, too, still needs to be tested.

Computing visual evidence. Ultimately all sources of localization information are combined and a decision is reached. Of the various glimpses in which the test flash is represented, the automatic glimpse triggered by the test flash itself will be the main determiner of localization, and this is illustrated in Fig. 9, Fig. 9a and b illustrates the weights of the pre-saccadic, motion, and post-saccadic events in the automatic glimpse. The test flash also may occur in the pre- and post-saccadic glimpses. Their weights in the overall computation, which are disregarded in this illustration, would be determined by the attentional strength of the test flash within these events.

A full computational model requires specification of the following major component processes (as well as many subsidiary ones), (a) It must specify the gating function that computes the precise weighting of events in the temporal neighborhood of the test flash. It has been assumed here that the gating function is similar to the automatic attentional glimpse of Weichselgartner and Sperling (1987). In these experiments, a visual test flash of an outline square was linked primarily to visual events within ± 50 ms of its occurrence (Fig. 8). (b) A formal specification is needed of the decision rule – how the weighted mixture of events linked to the test flash is interpreted to generate a response. For example, if a test-flash glimpse assigned 20% of the weight to the prefixation field, 45% to the motion field, and 35% to the post-fixation field, what localization response would be generated?

The simplest decision rule occurs when the link of the test flash to the motion segment is very weak, i.e., when the motion segment in Fig. 9b is very brief and its weight w_2 is negligible. This usually occurs in displays in which the test flash is far away in space or time from the nearest test marker spot. A reasonable initial hypothesis for this decision rule is that such a location judgement is binary – appropriate either to the pre- or to the post-fixation background, depending on which has greater strength. With a significant motion weight, intermediate location responses become more plausible. An intermediate location response would be generated, reflecting the relative strengths of the pre- and post-fixation backgrounds. However, at saccadic velocities, the motion is so fast that even when motion enters into the localization computation, it enters only as a weight to determine whether an intermediate response is appropriate and not as a cue to localization.

Close proximity of the test flash to a background marker yields more convincing, and hence more persistent, spatial evidence of test location than when the test flash is far away from the nearest marker. (The definition of near/far depends on distance from the fovea.) While this can be formally modeled as a distortion of spatial localization in the neighborhood of spatial markers, it represents an inherent complexity of vision. The perceived spatial location of a test flash during image motion depends critically on the nature of the moving image itself.

6.4.2. Model for visuo-motor localization during saccadic eye movements

The model for visuo-motor (extra-retinal, nonvisual) localization during real, rather than simulated, saccades is basically analogous to the model for purely visual localization. Instead of linking the test flash to three visual events, it is linked to three corresponding nonvisual events: (1) a representation of eye position relative to the head position before the saccadic movement; (2) a representation of the saccadic movement command itself (outflow) or the saccadic movement itself (inflow); and (3) a representation of eye position relative to head position after the movement. As in the purely visual model, the glimpse in which the visual test flash is contained assigns weights to each of these nonvisual representations; the relative strength of the link is determined by the degree of temporal overlap of the test flash with the nonvisual event. Indeed, in the nonvisual as in the purely visual computation, a simple weighting function which uses overlap with the motion segment only to determine whether or not intermediate location judgements might be justified, and then uses the ratio of strengths of post- to pre-fixation links to determine the location, might suffice to account for the data. The advantage of this kind of computation is that it offers good resolution of position of test flashes which occur during saccades without requiring any correspondingly fast visual processing. We will return to the issue later.

6.4.3. A model for resolving conflicting cues

A natural question to ask is: when both purely visual and nonvisual information are available, how are these sources of information utilized in the performance of various tasks? A linear model based on Thurstone's Case V (Thurstone, 1927) for the combination of perceptual cues has been found to work remarkably well in a variety of tasks involving visual judgements (Dosher et al., 1986; Bruno and Cutting, 1988). Predictions work equally well when cues agree (add) and when they conflict (subtract).

Weighing the evidence. In the case of two perceptual alternatives (such as two different perceptual interpretations of a rotating Necker cube), each perceptual alternative is represented by a pan on balance scale. Each cue represents evidence, and a weight proportional to the weight of its evidence in favor of an alternative is placed on the balance pan of that alternative. The relative importance of a particular cue depends not only on the abstract quality of information provided by the cue itself but also on how each subject weights that quality in the particular task. The final perceptual decision is determined by the algebraic addition of the strengths of all the cues plus a random error that reflects the variability of judgement (Fig. 9d).

For localization judgements during real saccades, the present evidence suggests that visual information is given greatest, perhaps exclusive, weight in making visual judgements (Hansen, 1979). For example, in localizing one visual event (a flash) relative to other visual events (the saccadically moved background), visual information appears to dominate. With increasing saccade size the situation may be different because two factors come into play. First, if the background markers against which a test flash is localized are not size scaled, they become less effective stimuli as they

become more peripheral. For this reason, saccade size would be best varied by varying the viewing distance and keeping the display constant. Otherwise, placing markers and test flashes more peripherally alters the early computations of visual spatial relations. Usually the alteration is in the direction of weakening the contribution of the spatial visual information. On the other hand, it seems obvious that nonvisual information would be stronger for larger saccades than for smaller saccades. Thus, even in purely visual localization judgements, nonvisual information may come to play a role in larger saccades. This relative increase in nonvisual influences on localization in large saccades does not involve a change in strategy (decision weights) but rather a change in the strengths of sensory inputs.

Motor localization reponses. In localizing targets by means of motor responses (versus making perceptual judgements) visuo-motor information is given greatest weight (versus purely visual information). For example, as was noted above, when an observer is asked to strike the location of a flash seen during a saccade with a hammer, the response is extremely accurate (Hansen and Skavenski, 1977, 1985). In this experiment, visual cues to flash localization were removed so that the observers were forced to rely on nonvisual information. Indeed, in localizing the flash relative to their body position, the observers did not succumb to the mislocalizations that they would have made if they were judging the flash's location relative to its visual environment. For further evidence that different sources of information are used in purely visual and visuo-motor saccadic localization tasks, see Ch. 5 of this volume.

6.5 *Motion perception during saccadic eye movements*

Suppose that, during a saccadic movement, the experimenter tricks the subject and shifts the visual field. To what extent can the subject detect such trickery? This question devolves into several component questions.

(1) When the retinal image sequence that would have been produced by a saccadic motion is artificially perturbed by extraneous motion, how detectable is the perturbation? (Motion masking.)
(2) How well can the observer detect that his eye has not landed where it intended? (Saccadic calibration.)
(3) Given that motion is a perceptual primitive, why is image motion not experienced during saccades? (Motion suppression.)

Motion detection and discrimination during saccades. Earlier, this chapter considered the detection of a simple flash during a saccade. Early investigators had claimed there was great loss of sensitivity during a saccade. Once the proper control experiments were conducted with simulated saccades, it became clear that the actual sequence of images on the retina produced visual masking and positional uncertainty of the test stimulus that was sufficient to account for threshold changes observed during saccades. There was no residual threshold change that could be attributed to the saccade per se. The problem of detecting and discriminating visual motion during saccades seems to be similar. There are several reports of an impaired ability to detect or discriminate motion during saccades (e.g. Bridgeman et al., 1975; Mack, 1970, Stark et al., 1976; Whipple and Wallach, 1978). However, when Brooks and her collaborators (Brooks and Fuchs, 1975; Brooks et al., 1980a,b; Brooks and Impelman, 1981) produced equivalent motion perturbations in real and in simulated saccades, they found them to be equally detectable.* Their answer to the motion masking question raised above is that any inability to detect motion perturbations during eye movements is explained entirely by the sequence of images on the retina. To this it must be added that in both real and simulated saccades, Brooks et al.'s subjects discriminated normal from perturbed motion on the basis of the shape of the perceived motion blur – not on the basis of perceived motion of the perturbation (Brooks, personal comunication).

In Brooks et al. (1980b) and in every other instance up to this point, when a psychophysical discrimination was based on retinal images, it has not mattered whether these images were viewed in a stationary or saccadically moving eye provided that the retinal images were equivalent. However, the problem of *motion* detection and discrimination during saccades is more complex than the problem of simple object detection because the *sensation of image motion obviously is suppressed during saccades.*

The sensation of motion: saccadic motion suppression. Consider Question 3: saccadic trajectories are excellent stimuli for the motion perception system. That is, moving an image in the trajectory of a simulated saccade on a stationary retina produces a strong sensation of apparent motion.* Why do we not experience visual apparent motion during the same retinal image movement when it is saccadically induced? Notice that the sensation question is a fundamentally different question from the discrimination question raised above. Sensation refers to how an observer describes his experience, and does not involve a right or wrong answer as does a discrimination task.

It is helpful to place the suppression of motion sensations during saccades into the broader context of other voluntary movements. For example, the otolith signals acceleration with respect to gravity: why don't we experience a sensation of falling whenever we voluntarily sit down? When we voluntarily turn our head, the vestibular and visual systems should signal vertigo, but we do not experience it. The other side of the coin is the shock we experience when we expect one sensory input, for example, strawberry mousse, and encounter another, sour cream with salmon roe. However, merely observing that the interpretation of sensory input quite generally depends on voluntary movements and on the corresponding sensory expectations does not answer any specific questions about the processes that are invoked. We first consider Question 2 (Saccadic calibration) and then the issues concerning the types and levels of sensory processing.

Saccadic calibration: piano analogy. The answer to Question 2 concerning sensitivity to environmental displacement during saccades depends on the size of the displacement as a fraction of intended eye movement, and on the suspicions of the observer. Sperling and Speelman (1965) observed that a stimulus displacement of 2 deg during a 4-deg saccade was reliably detected (cf. Bridgeman et al., 1975; Stark et al., 1976; Whipple and Wallach, 1978).

The issue of sensitivity to visual displacements can be easily understood by an analogy. Imagine a pianist performing a difficult piece on a piano. While his hands are in the air, we move the piano. Indeed, on grand pianos, the soft pedal accomplishes just such a movement, moving the keyboard by about a quarter the width of a key so that the hammers strike only two of three strings. Such small keyboard movements usually go unnoticed. Suppose, instead, that while the pianist's hands are in the air, we move the piano the width of a key. The unfortunate pianist would strike wrong notes and probably infer that he was out of practice. But, suppose we moved the piano a foot, so that the pianist's hands struck completely unbelievable notes. Not only would such trickery be shocking and instantly recognized, but the pianist would attribute every subsequent wrong note to external interference. In a psychophysical procedure, the pianist's ability to discriminate real mistakes from induced mistakes would follow roughly a Weber fraction of the lateral hand movement. The situa-

* While Brooks et al (1980a) mostly found identical thresholds in real and simulated saccades, in a few of their conditions there were slight differences that could have been caused by residual, uncontrolled differences in procedure.

* The motion perception of objects moving at saccadic speeds has been studied almost exclusively with part-field stimuli. Saccades move the entire retinal field and, except in the laboratory, the entire retinal field is usually filled with stimuli. The author is not aware of any study that examines whether full field stimuli moving at saccadic velocities produce as strong a perception of motion as part-field stimuli.

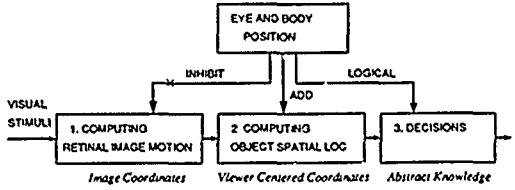


Fig 10 Processing of visual and nonvisual information. Three stages at which nonvisual information interacts with visual information the label on the arrows indicates the mode of interaction

tion with saccadic eye movements is not essentially different Whether we move the hand or the eye, we know what to expect when we arrive. The cognitive processes by which we build up representations of the external world and derive predictions from the representations are quite complex; they are considered briefly later in this chapter.

The experiments on the spatial localization of flashes during eye movements are analogous to the localization of tactual stimulation during ballistic hand movements. As with the eye, when the hand is stimulated as it is moved rapidly over a surface, the source of stimulation is projected onto a location on the surface. The subject's hand trajectory can be compared quantitatively to the objective trajectory, just as with eye trajectories, and it is undoubtedly subject to similar illusions. Indeed, comparisons of motor programming in vision with motor programming in other modalities promise to yield insights into both domains of study

Determining the level of visual–nonvisual interaction in the perception of image motion To analyse the processing of visual motion during saccades, we consider the visual and nonvisual inputs and the three levels at which they can interact. Specifically, the inputs are (1) the retinal images generated by the saccades and (2) the nonvisual eye movement and body position signals. The computational levels are (1) computing retinotopic image motion, (2) computing object spatial position, and (3) a decision level (Fig. 10). At each level, we consider computations that could inhibit the sensation of motion during saccades.

At the level of computing retinotopic motion, the effect of nonvisual input from a saccade would have to be inhibitory and nonspecific. That is, because all parts of the visual field may be stimulated by saccadic motion, motion signals would have to be suppressed throughout the visual field The question of whether motion signals that represent directions of motion counter to or perpendicular to the saccade are also suppressed (Stark et al., 1976; Whipple and Wallach, 1978) is left open because, it will be argued, the retinotopic inhibitory mechanism is itself implausible.

A higher level of saccadic visual–nonvisual interaction is at the level of computing the position of a visual object relative to the head. (For specificity, we take head direction to be the direction the nose is pointing.) Computing object position requires adding two angles: (1) the retinal angle between the object and the fixation point – the line of regard – and (2) the angle between the line of regard and the nose. Saccadic motion suppression at this level implies that spatial position rather than the retinotopic motion is used to infer perceived motion.

At the decision level, a decision is made about whether or not object motion may have occurred during a saccade Consider the piano analogy. The pianist's hands land on the piano but on a wrong note Ordinarily, the pianist does not entertain the hypothesis that the piano has moved, and so the sensory signal is logically re-interpreted to indicate that the hands must have erred in executing their intended movement. The case of visual objects whose position is perturbed during a saccade is similar. All the systems up to the point of decision may be sending appropriate signals, but they are discarded at the point of decision. However, when the new game is pointed out to subject or pianist, the decision rule can be quickly revised.

Modifiability, the critical role of feedback. It is assumed here that the modifiability of processing is related to level: the higher the level, the more easily processing is modified.

Full-field inhibition of retinotopic motion computations during saccades would generally be assumed to be an unmodifiable genetically determined process. At the level of computing coordinates, the ability to calibrate eye movements is genetic, but saccadic extent (and many other motor components of eye movements) can be recalibrated in a few minutes to a few hours of observation time (Keller and Zee, 1986). The situation is not essentially different from the case of a pianist switching from a standard piano to a harpsichord which has narrower keys or, in the second case, having different dynamic properties so that key-press movements have to be recalibrated.

The issue of modifiability is critically related to experimental method. Feedback means that a subject is told and/or experiences the consequences of correct versus incorrect responses. When they are correctly carried out, experiments without feedback are essentially ecological investigations; they determine how sensory inputs are habitually computed. Experiments with full feedback can determine the computational limits. For example, to establish that there is retinotopic motion inhibition requires an experiment with feedback. From experiments without feedback one can learn only that subjects habitually ignore motion signals during saccades. To determine that the failure to perceive motion during saccades is not merely a habit but an unmodifiable deficiency would require the experiment with feedback to fail to train subjects to use retinotopic motion signals.

To understand the limits of motion perception during saccades requires at least two conditions: (1) comparing real saccades with simulated saccadic controls and (2) experiments with feedback. The numerous reports of saccadic motion suppression cited above fail one or both of these criteria. We consider saccadic motion perception below in the section on correlating successive saccadic images.

The issues suggested by the flow diagram in Fig. 10 are, in principle, resolvable. For example, a possible generalized loss of sensitivity to image motion during saccades can be investigated experimentally by presenting visual motion stimuli (motion probes) before, during, and after saccades. There are the masking and suppression questions: the ability to experience the perception of motion of the probe and the ability to discriminate different probe motions. Generality of suppression is studied by determining to what extent perception and discrimination of motion are inhibited by temporal proximity of the motion probe to a saccade, and determining whether the direction of motion matters. The problems in pursuing this research are great technical requirements in eye movement recording and even greater technical difficulties in producing the proper simulated-saccade controls.

One argument against motion suppression occurring early in visual processing is the autokinetic effect. Stationary points of light in the dark, even when fixated, appear to make small movements from time to time (see Mack, 1986, for references). Autokinesis seems like an obvious failure of motion suppression for small eye movements, although it is not unexpected, given that nonvisual information appears to be quite weak for small saccades. On the other hand, with normal full field stimulation in the light, the world as a whole does not appear to jump around, suggesting that more stringent perceptual criteria are applied to large-field than to small-field motions. In summary, the question of "At what processing level, and by what mechanisms are the sensations of motion re-interpreted during saccades?" remains unresolved.

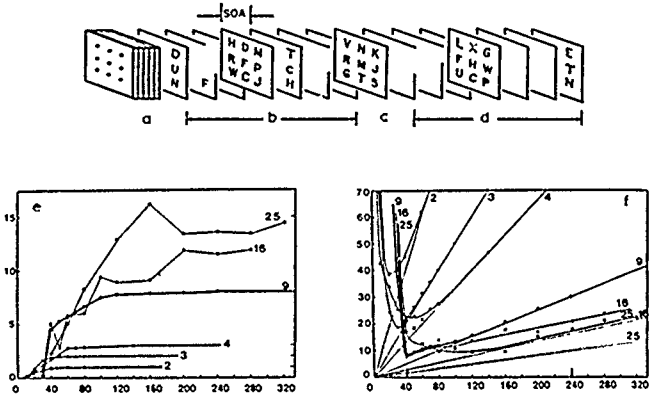


Fig 11 Simulated saccadic search. A fixation field (a) is followed by a variable number (b) of arrays which contain letter distractors. The critical array (c) is embedded near the middle of the sequence It contains a numeral target, and is followed by (d) letter arrays The time between successive arrays is SOA (stimulus onset asynchrony) (e) The estimated number of characters in each array successfully searched by the subject as a function of the SOA Data for the 9-character array illustrated above are emphasized by a heavy connecting line The lack of improvement with SOAs longer than about 120 ms for arrays of 16 and 25 ms indicates that the second half of the exposure interval is wasted when the interval is as long as the intersaccadic interval (f) The same data replotted to show the estimated scan time per character 40 ms indicates the quickest scan times for 9- and 16-letter arrays, 240 ms indicates the minimum intersaccadic interval (Redrawn from Sperling et al., 1971. Data from subject *J.S.*)

7. Sequences of saccades

7.1 Visual search

7.1.1 Simulated search: saccades are not always the optimum information gathering strategy

By the definition of 'survival of the fittest', the visual system that we now have, which includes a high-resolution fovea, a low-resolution periphery and saccadic eye movements, is the optimum visual system under the set of constraints under which humans evolved. In this light, it is informative to consider a simple search task in which the saccadic mechanism is nonoptimal because it is too slow.

Consider a simple search task such as scanning a large array of letters (distractors) for an embedded numeral (the target) Normally, this search would be carried out by means of saccades, each saccade positioning the eye in a different region of the array To simulate this search task in the stationary eye, the new information is presented to the stationary eye at regular intervals, as illustrated in Fig. 11. The subject views a long sequence of stimulus frames, all except the critical frame contain arrays of letters In the critical frame, a number is embedded at an unknown position of the array. The subject's task is to detect the number. As a purely technical matter, the subject is also required to press a reaction-time key as quickly as possible upon detection, to give the position in the array at which the number occurred, and to state a confidence for the identification task. These additional bits of information are useful in discriminating true from accidentally correct detections.

The results of such an experiment are the percentages of correct detections as a function of the parameters of the experiment (Fig. 11e). The percentage of correct detections is most usefully transformed into a search rate – the number of distractors that must be searched each second in order to support the observed percent correct (Fig. 11f). The parameters include the time interval from one array to the next (stimulus onset asynchrony, SOA), the size of the arrays to be searched, the sets of items used as targets and distractors, the size and discriminability of the items, the advance knowledge that the subject may have about possible targets, and so on (Sperling and Dosher, 1986).

The parameter of interest with respect to simulated eye-movement search is the interval between arrays. Saccades do not occur faster than about 4/s; this corresponds to a presentation time of 250 ms per array in the simulated search procedure. Indeed, when arrays of characters are presented every 240 ms, the observed search rate is about 50 characters per second, which is equivalent to 20 ms per character (Sperling et al., 1971). About the same search rate is observed in natural search (Neisser, 1963, 1964; Neisser et al , 1963). However, when the interval between arrays is reduced, the search rate can be substantially higher. The highest rate of search, more than 100 characters per second (less than 10 ms per character), occurs when new arrays occur every 40–50 ms, a presentation rate five times faster than the rate of eye movements. The simple empirical conclusion is that eye movements, which limit the time between bursts of new information to one per 250 ms, limit the rate of search to half of what it can be when the presentation rate is increased.

Why is this particular search task (a numeral among letters) slower with natural eye movements than in the simulated search procedure? According to analyses by Fisher (1982) and Sperling and Dosher (1986) there are two interlocked reasons. First, the visual system seems to be able to execute the search in parallel in at least three of four locations of the visual field. Second, foveal search is faster than peripheral search. To some extent, foveal/peripheral search differences can be overcome by appropriate size scaling of stimuli to be searched. However, mixing character sizes in arrays to be searched slows search down rather than speeding it up (Sperling and Melchner, 1978). Therefore, the fastest search occurs when arrays of four or more characters are presented to the fovea at a rate consistent with its information-processing capacity. For highly legible characters, central vision can process 25 batches of four characters per second. This mode is two times more efficient (in terms of the number of characters searched) than saccadically driven search, which processes only 3 or 4 batches per second.

7.1.2. When is saccade rate a limiting factor in performance?

There is enormous flexibility in visual processing The only task in which saccades have been a limiting factor was the simulated search for a relatively large, highly familiar target (a numeral) among letter distractors (Fig. 11). This is contrary to intuition, which suggests that saccades would limit performance when searching for tiny targets which could be discriminated only in the fovea. The problem is that, when a target is made so difficult to detect that it requires foveal acuity, the processing time to detect that target, even in the fovea, is likely to become so long that processing time itself, rather than intersaccade time, becomes the limiting factor. All this merely indicates that the capacities of the motor and processing components of the visual system are matched to each other, which is as it should be.

7.2. Do saccades initiate processing episodes? The optimal duration of inter-saccadic fixations

When the retinal image is artificially kept motion-

less, independent of eye movements, it fades, and information-processing from that image ceases (Ditchburn and Ginsborg, 1952, Riggs et al, 1953; Yarbus, 1957). Saccades can maintain or restore an image to visibility; however, they are not necessary – smooth eye movements are perhaps even more effective (e.g., Gernts and Vendrik, 1974; see Kowler and Steinman, 1980, for a review). These observations have induced vision scientists to speculate that smooth eye movements are especially associated with continuous information-processing whereas saccades are associated with discontinuous, episodic information-processing. Saccades are assumed to initiate processing episodes. The visual system is assumed to be especially adapted to process the kinds of image sequences that saccades provide: tens of milliseconds of smear followed by hundreds of milliseconds of steady image, repeated over and over. Undoubtedly the visual system is adapted to this saccadic mode of operation. This chapter continues, with more complex stimuli and tasks than before, to deal with the question raised at the onset: to what extent do the information-processing adaptations to saccadic vision operate independently of the saccades themselves? In particular, to what extent are adaptations to saccadic vision exhibited equally when the simulated saccadic sequence of images is presented to a stationary eye?

7.2.1 In natural saccadic viewing, are long-duration fixations 'better' than brief ones?

Loftus (1972) investigated this question experimentally using natural saccadic movements in a recognition memory experiment. In the learning [illegible], subjects viewed a sequence of natural scenes, each scene [illegible] for a fixed time interval. Later, subjects were tested for their ability to discriminate previously viewed scenes from new (distractor) scenes. Loftus found that the best predictor of later recall was the number of saccades that a subject made in the initial viewing. Exposure duration itself had an effect upon recall only in controlling the number of saccades: given the same number of saccades, the inter-saccadic durations themselves had no influence on recall.

Loftus's finding that increasing inter-saccadic duration has no effect on recognition accuracy invites the inference that each saccade initiates a processing episode which is completed in less time than the shortest inter-saccadic duration. The possible difficulties with this conclusion illustrate the problem of studying saccades naturally without also using an appropriate simulated-saccade control. The problem is that saccadic viewing strategies are determined by the subject, not the experimenter. Therefore the viewing strategy may be perfectly confounded with the intrinsic memorability of a picture. Easy-to-remember pictures induce short inter-saccadic durations. Without further embellishments, Loftus's procedure would admit no conclusions about the effectiveness of saccades as a function of the inter-saccadic durations. This kind of difficulty in studying natural saccadic viewing is very difficult to overcome because inter-saccadic duration is a dependent variable rather than an independent variable. Loftus himself ultimately found it necessary to study approximately-simulated saccades (Loftus, 1981). Sometimes, it is desirable, additionally, to use artificially constructed stimulus materials to give still better experimental control, as in the attempt to answer the following question.

7.2.2 In a search task, are two short saccades better than a long one?

The question of whether saccades initiate processing episodes that are quickly over – even before the onset of the next saccade – suggests several experiments with simulated saccades. For example, in a simulated search task, are two saccades better than one long saccade? And, if two short saccades are indeed better than one long one, must the information presented in the two successive exposures fall on different retinal coordinates?

Letter arrays Kowler and Sperling (1980) studied simulated search for a numeral embedded in a 5×5 letter array viewed with either single or double exposures of various durations. Each stimulus sequence was terminated with exposure of a visual noise field. The stimulus array was either flashed once (a), or twice (b), or twice with a lateral translation between exposures (c), or the array was presented continuously (d), or continuously with a lateral translation in mid-exposure (e). Search accuracy depended little on the viewing condition when the total duration of visual availability (onset of the stimulus to onset of the noise field) was less than about 100 ms. For longer exposures, the order of conditions, from best to worst, was d > e > b > c > a. That is, displacements were not helpful, two flashes were better than one, but a single continuous exposure was always best. Even for the longest simulated fixations (800 ms), dividing the long fixation into two short ones was harmful under the parameters of this search task.

Natural scenes To complement his study of natural scenes viewed by natural eye movements, Loftus (1981) studied natural scenes viewed by successive bursts of illumination, each burst followed by a visual noise field. His procedure was not a saccadic simulation because the time interval between successive exposures was long enough to permit real saccades to reposition the eyes. Like Kowler and Sperling (1980), Loftus (1981) found that breaking a long flash into several shorter ones did not improve recognition memory. However, when fixation changed voluntarily between flashes, performance did improve with the number of fixations. Loftus concluded that the critical component in later recognition is the number of picture-features that a subject remembers from a scene. For example, generally, performance improves with number of flashes. But, "when the number of places looked is held constant, the effect of number of flashes vanishes, thereby indicating that additional flashes are only useful insofar as they permit acquisition of information from additional portions of the picture" (Loftus, 1981, p. 373). Memorable features in Loftus's natural scenes were less dense and more widely spread out than were characters to be searched in Kowler and Sperling's 5×5 arrays, so subjects benefited from successive images that fell on different retinal locations in the scene experiment and not in the character search experiment.

7.2.3. Are sudden onsets (such as might be provided by saccadic eye movements) necessary or beneficial for information-processing?

To directly test the utility of abrupt stimulus onsets for information-processing, Kowler and Sperling (1983) used a simulated saccadic sequence of images in a search task for a numeral embedded in a sequence letter arrays, as shown in the top panel of Fig. 12. Additionally, in various conditions of their experiment, the temporal waveform of the successive images was varied. They measured search accuracy with both abrupt (step) and gradual (sawtooth) onsets and offsets of images in the sequence, at two presentation rates (Fig. 12).

Search accuracy was the same, independent of the waveforms shown in Fig 12; search accuracy depended only on the time between successive arrays. That is, only the time available to process the stimulus items influenced performance, not how that time was apportioned into dark and light phases of the cycle. These results are quite different from those obtained with stimuli at the threshold of detection or discrimination. When visual process-

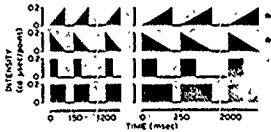


Fig 12 Sequences of stimuli whose intensity waveform is varied according to four different functions at each of two presentation rates. Each triangular (or square) packet represents the exposure of a new array of characters in a search experiment. Because ramp-on stimuli do not have sudden onsets and all the other stimuli do, a theory which asserts that sudden onsets initiate periods of information-processing would predict (incorrectly) that search performance is inferior with ramp-on stimuli. (From Kowler and Sperling, 1983.)

ing is limited by the energy in the stimulus, the temporal waveform of the stimulus matters critically for performance (Watson and Nachmias, 1979). However, when stimuli contain sufficient energy to be easily discriminated processing is time-limited, not energy-limited, and the temporal waveform becomes relatively unimportant (Kowler and Sperling, 1983, Sperling, 1979).

The conclusions are that, under good visibility conditions, the physical parameters of image onsets imposed by saccades are relatively unimportant, presumably because of the great efficiency of visual preprocessing. For a particular stimulus and a particular task, high-level processes of feature encoding (in recognition memory experiments) and of feature matching (in search experiments) determine where the eye should be placed and when it should be moved for optimal performance.

7.3 Two-flash displays: masking, localization, movement, memory

The two-flash paradigm. Experiments with a two-flash stimulus have been particularly productive in the analysis of information-processing within and between fixations. The observer views two consecutive, brief flashes separated by a time interval δt. Either both flashes are confined to within a single fixation (the Within condition, Fig. 13) or a saccadic eye movement occurs between the flashes (the Between condition). Comparison of performance in Within and Between fixation conditions yields insight into saccadic information-processing. The ability of the observer to correlate the contents of the two images is tested by memory tests or by psychophysical tests that involve, for example, the ability to perceive motion between the images. The previous section considered search experiments in the within-fixation but not the between-fixation variant of the paradigm. The two-flash paradigm has also been used to study spatial localization and visual masking.

The great technical advantage of the two-flash procedure is that the simulated eye movement control experiment does not require producing a complex eye movement streak on the retina, merely two flashes. As with other eye movement paradigms, there are two disadvantages of the two-flash paradigm. The exact time of occurrence of a saccadic eye movement cannot yet be perfectly predicted. Many trials must be conducted in order to obtain a few critical trials with the eye movement centered between the flashes. To control for possible effects of the context of imperfect trials on performance in the critical trials, an equivalent imperfect context has to be provided for the within-fixation control experiment. The second problem is that the eye position must be known exactly at the time of the two flashes in order to ensure that the control presentation in the within-fixation presentation is truly equivalent to the between-fixation presentation – for both the 'same-retinal-coordinates' and 'same-spatial-coordinates' variations. While the problem of positional accuracy is endemic to all eye movement recordings, the apparent simplicity of the two-flash paradigm has seduced experimenters into attempting it with less-than-adequate eye movement recording.

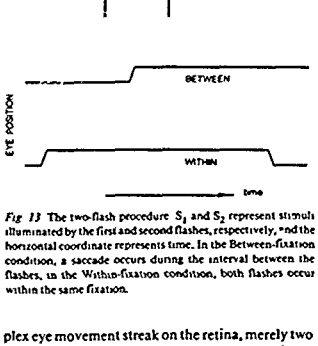


Fig. 13 The two-flash procedure. S_1 and S_2 represent stimuli illuminated by the first and second flashes, respectively, and the horizontal coordinate represents time. In the Between-fixation condition, a saccade occurs during the interval between the flashes, in the Within-fixation condition, both flashes occur within the same fixation.

7.3.1. Visual masking

Visual masking in a two-flash paradigm with an interleaved saccade was studied by Davidson et al. (1973) and Irwin et al. (1988), with roughly similar procedures and results. In Irwin et al.'s experiment, subjects were presented first with a 10 ms exposure of a row of five letters. This was followed by a 40–70 ms blank interval during which, on some trials, a saccade occurred. After the saccade, a masking pattern was superimposed on one of the letters. The masking pattern was found to exert its masking effect primarily when it occurred at the same retinal location, not the same spatial location, as the letters.

Does interposing a saccade between the first and second flash alter the masking effect of the masking pattern? With a δt between onsets of the first and second flashes of 40–70 ms, retinotopic masking results suggest that an eye movement would make no difference for this kind of masking. However, neither Davidson et al. (1973) nor Irwin et al. (1988) report a no-movement control condition, so we can only conjecture that masking is the same in real and simulated saccades.

7.3.2 Spatial localization

To compare masking and spatial localization during saccades, Irwin et al. (1988) used a two-flash background presentation like that described above. Again, the first flash was a 5-letter array; the second flash was a bar marker rather than a masking field. The bar marker was a short vertical line segment which instructed the subject to report the name of the letter below it (Averbach and Sperling, 1961). These experiments are essentially spatial localization experiments completely analogous in many details of procedure, theory and results to those described in the section on spatial localization during saccades (section 6.2). Irwin et al.'s results show that the bar marker changes its apparent location relative to the stable letter array with approximately the same time course during the saccade as did Sperling and Speelman's (1965) short-line segment relative to their dot array in the localization experiments described earlier. The comparison between experiments is only approximate because Irwin et al. did not make precise measurements of the time course.

Irwin et al. observed that, even though a masking flash that masks an earlier letter was at the same retinotopic location, the apparent spatial location of the masking pattern corresponds to its new spatiotopic retinal coordinates. That is, the apparent spatial location of a masking pattern is computed in the same way as the apparent location of other spatial patterns, such as bar markers.

7.3.3 Motion detection and perception

The two-flash paradigm naturally lends itself to the study of motion perception between the two flashes. Perceiving motion requires some form of correlation to be computed between the first and second stimulus, so motion perception implies at least an elementary form of pattern memory. The issues that arise in the two-flash paradigms are precisely the same as those which emerged in section 6.5. However, in experiments that measured the ability of subjects to detect object displacements during saccades, no attempt was made to ascertain whether detection was based on perceived motion or on perceived change in location.

Shioiri and Cavanagh (1989) attempted to determine whether motion could be perceived during a saccade by using random-dot patterns which offered good motion cues but only weak locational cues when they were displaced. When pattern displacements occurred between two fixations in an explicit two-flash paradigm, their subjects failed to discriminate displacement from no-displacement trials. The subjects were also unable to use apparent motion to correctly identify the direction of displacements that occurred around the time of saccades.

Unfortunately, Shioiri and Cavanagh's procedures illustrate the hazards of violating the three methodological precepts proposed above. They did not measure eye movements accurately enough to know the actual retinal placement of their stimuli. Perhaps for this reason, they did not use feedback to teach the observers to use all the available move-

ment information. Therefore, the most we can know is that their observers habitually do not use retinal movement information to determine whether stimuli have moved during saccades – not that motion information is unavailable or suppressed. And the investigators did not run the simulated-movement control experiment within a fixation to permit comparisons of movement perception in retinally matched sequences of displays within and between saccade-separated fixations. Thus, while it is clear that people tend not to report perceiving motion between two saccade-separated flashes, the question "To what extent *can* motion be perceived between two saccade-separated fixations?" remains unanswered.

7.3.4 Recognition memory for images related by translations

While motion perception is an elementary computation that compares two (or more) views of the world, there may well be analogous higher-level computations. Consider that saccadic eye movements convert the visual input into essentially a series of still frames at a typical rate of about two or three frames per second. When the environment is stationary, all these successive images are related by simple translation. Might there be a specialized memory for recognizing and storing images which differ only by translation? *How are the relationships between images coded to enable the observer to build up a coherent internal model of the world.*

Recognition memory for translated images in the stationary eye Let two successive images, such as might be produced by successive saccades, be produced on the stationary eye. Does the observer have any special ability to recognize relationships between such successive images? In one procedure, Spelman and I presented subjects with successive images, each consisting of ten shapes. One shape was changed, the remaining ones were the same in both presentations. The subject's memory was tested by asking which shape was different. Shapes were chosen that were not as easily named as alphanumeric characters, and brief exposure durations were used in order to selectively probe visual memory (and to reduce the role of verbal memory). The ten shapes were arranged in a horizontal row, and translated upward or downward. The direction of translation (up or down) and the distance between frames was varied randomly.

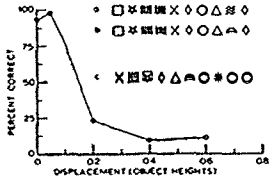


Fig 14 Stimuli and results from a two-flash, simulated saccade memory experiment. The first stimulus is *a*, the second stimulus is *b*; the subject's task is to locate the changed position. Stimuli *a* and *b* occur in quick succession separated vertically, as shown. (c) Two stimuli that overlap by about 0.2 object heights. This small shift reduced performance from near perfect to less than 25% correct. The graph shows recognition accuracy (identification of changed locations) from an experiment in which displacement between *memory* and test stimuli was varied. The data points are the average of three subjects, the accuracies have been corrected for chance guessing. (Sperling and Speelman, unpublished data, AT&T Bell Labs, 1968.)

Fig. 14 shows typical results of the study in which the second flash immediately followed the first and the duration of each was 167 ms. When the shapes moved, strong motion cues were produced by the movement. In addition, the changed target item produced dramatically more focal flicker and random apparent motions. The combined flicker/motion cue allowed the subjects *easily to identify the* location of the changed shape. However, the subjects had no *special ability to identify* the first of the two shapes in a location. Virtually identical results were obtained when shapes were moved horizontally or diagonally, when they were viewed from different distances, and when the subject knew in advance the probable direction of movement. There is a potential technical problem in studying large displacements which place the two rows one above the other, because visual persistence of the first array facilitates comparison with the second. However, under conditions that minimize visual persistence or when post-stimulus masks are used (Sperling, 1963), performance continues to decline monotonically with the magnitude of displacement regardless of its direction.

When the shapes moved *for 1/5 of their extent or* more, or there was a sufficiently long interstimulus *interval so that motion* cues did not selectively point to the altered shape, subjects did quite poorly. Subject's asymptotic performance can be characterized as indicating a memory of slightly more than one of the ten shapes. If the altered stimulus was the stimulus they had memorized, they detected the change; if not, they guessed randomly. There was no indication whatever of a special memory for translated images.

7.3.5. Trans-saccadic perceptual fusion

The perceptual fusion paradigm. Perceptual fusion *refers to the phenomenon* in which two consecutive displays are perceptually combined and perceived as a unitary display. For example, in the two-flash display of Eriksen and Collins (1967), some subareas of a stimulus letter are displayed in the first flash and the remaining ones in the second flash. When the first and second flashes occur in extremely close succession, the resulting stimulus is not discriminably different from a single exposure of the whole stimulus, and the letter is clearly identifiable. As the time between flashes is increased, perceptual *fusion* becomes increasingly difficult and, at around 100 ms of separation, depending somewhat on stimulus conditions, accuracy of letter identification drops to chance. In the two-flash stimulus developed by Hogben and DiLollo (1974), the first flash contains 12 dots randomly chosen from a 5×5 square array; the second flash contains 12 of the remaining 13 dots; and the subject's task is to locate the missing dot. Again, when the two flashes are presented in extremely close succession, the subject perceives 24 dots simultaneously, and the location of the missing dot is found effortlessly.

As the interval between flashes is increased, performance eventually drops to chance levels. How does interposing a saccade between the two flashes affect perceptual fusion?

Fusion requires retinal, not spatial, superposition An early study (Jonides et al., 1982) of perceptual fusion in the missing-dot paradigm erroneously reported that when flash 1 was presented in the periphery before a saccade, and flash 2 in the fovea after the saccade, there was good perceptual fusion. That is, the two flashes in the same physical location, but different retinal locations, could be combined to solve the missing-dot problem. Subsequently, this result was discovered to be an artifact of luminous persistence in the CRT display (Jonides et al., 1983).*

With correctly constructed displays, there is not more perceptual fusion in the between-fixation (interposed saccade) condition than in the within-fixation control. This was demonstrated in a letter-fusion paradigm by O'Regan and Levy-Schoen (1983) and in missing-dot paradigms by Bridgeman and Mayer (1983), Irwin et al. (1983) and Rayner and Pollatsek (1983). With an interposed saccade, two flashes that originate at the same physical location strike different retinal locations, but perceptually they seem to have originated from the same physical location. In the control condition in the stationary eye, flashes that fall on the same two retinal locations appear to have occurred at quite different physical locations. Each of these perceptions is correct. However, correctly perceiving retinally separated flashes to have occurred in the same spatial location does not imply useful trans-saccadic perceptual fusion.

Can perceptual fusion occur when two flashes strike the same retinal location but a saccade has intervened so that they appear to have occurred at different spatial locations? This question is quite similar to the two-flash motion question posed in

* Jonides et al. (1982) are not the only investigators to have erroneously reported trans-retinal perceptual fusion. See Irwin et al. (1983) for a critique.

an earlier section. In both cases (achieving perceptual fusion, detecting small retinal displacements) to succeed in the trans-saccadic task, the subject must succeed in ignoring or cancelling the non-visual signals arising from the saccade. And because in both cases there are formidable technical and procedural difficulties in conducting the experiments, we do not yet have adequate answers.

7.3 6. Other tests of trans-saccadic memory: conclusion

Among the contexts in which the notion of a special trans-saccadic memory has been proposed is reading (McConkie and Rayner, 1976). Here, too, experimental attempts to demonstrate such a specific memory have failed (McConkie and Zola, 1979; Rayner et al., 1980; McConkie et al., 1982). Psychophysical discriminations which require memory for line length and for the shape of rectangles demonstrate that there is trans-saccadic memory (Palmer and Ames, 1989). However, that subjects remember length or shape from one fixation to the next is hardly a novel discovery. The particular issue that concerns us here is whether a saccade, as compared to a simulated saccadic display in the fixated eye, facilitates or inhibits performance in the memory task. The reading and psychophysical experiments have not been designed to answer this question.

The two-flash experiments have not yielded any data to suggest that interposing a saccade facilitates performance relative to the simulated saccade in the stationary eye. Indeed, there remains the as yet unproved possibility that when a task requires the subject to ignore nonvisual signals generated by a saccade, performance may suffer relative to the simulated saccade.

7.4 Organizing information from sequences of saccades

7.4 1. Spatially coordinating successive retinal images

A failure to coordinate successive images Subjects with extreme tunnel vision are unable to coordinate the information from successive eye movements. A similar failure to coordinate images in the stationary eye was found by Hochberg (1968). Subjects viewed a sequence of frames that represented successive views of a complex shape. The subjects were unable to deduce the overall shape from these views. Apparently, the information about relative locations of points in successive views is not easily derived from a sequence of images.

7.4 2. Both image content and spatial location are represented symbolically

Spatial location as a tag In the retina and visual cortex, spatial location is coded retinotopically and anatomically. It is taken as axiomatic that at higher levels of processing, spatial location is ultimately coded as a tag, not as an anatomical brain location. That is, successive views are stored not in a topological arrangement corresponding to their two-dimensional relationships in the environment, but in more complex symbolic form in which information about relative positions in space is carried as a tag or feature in the representation. A visual object is described by a set of visual features and the relationships between them.* The representation of spatial location of the object relative to other objects, to the body, and to the environment is not logically different from the representation of other relationships. In this respect, the representation of visual space is not essentially different from the representation of tactual space defined, for example, by the hands moving over a surface and attempting to learn about it.

Must spatial tags be derived from eye and body movements? One interesting question is whether the information about spatial position that is derived from the position of the body, head and eyes can be replaced with position information derived from other modalities. For example, can a representation of the environment be built up when the information about location is provided by the position of the hand rather than the eye. To investigate this question, the eye is fixated on a display screen. The subject places a finger at various spots on a surface and an image is produced corresponding to the neighborhood of each spot pointed at. To the extent that the subject can learn to substitute finger movements for eye movements, the visual processing of successive saccadically produced visual images is not uniquely linked to the oculomotor system but can utilize other channels which provide reliable spatial information.

* See Ballard (1987) and Feldman (1985) for discussion of the frames of reference within which observer-object relationships are best represented.

8. Summary and conclusions

Smooth and saccadic eye movements are uniquely adapted to acquire information via an eye that is organized into a specialized fovea and a wide periphery. The most useful working hypothesis is that, while both visual sensory processes and motor control have evolved to a high degree of specialization to deal with the eye movements, modality-specific processes yield to content-specific processes as early as practicable in the processing hierarchy. Thus, in processing information acquired by pursuit and saccadic eye movements, the earliest link between the retinal and extra-retinal components of the eye movement appears to occur at high levels of processing. As a consequence, when differences on psychophysical tests of perception in responses to self-produced image motion and to imposed image motion were observed they were attributed, for the most part, to failures to provide truly equivalent retinal stimuli in the moving and the stationary eyes. For example, acuity seems to be determined by retinal slip, and it makes no difference whether the object or the eye is moving. Similarly, the visual system is designed so that the kind of motion smear produced during saccades is not perceived even when it is produced on a stationary retina by a simulated saccade. The many reports of changes of visual sensitivity during saccades are adequately explained by the masking effect of the actual sequence of stimuli on the retina and by the uncertainty in where a test stimulus that is flashed during a saccade will appear to be located.

Spatial localization of flash seen during a 4-degree saccadic eye movement did not differ from localization of a flash during the equivalent-imposed image movement. Errors of localization could be explained by assuming that there was a temporal uncertainty of about 6 ms in when a visual test flash occurs relative to saccadically produced image movement. Additionally, for some subjects, the subjective duration of their saccadic image movement was somewhat longer than its objective duration, and the subjective movement began too soon (relative to a test flash). This slightly inaccurate internal representation of the imposed image movement produced characteristic localization errors. For larger saccades, there were significant differences between localization judgements in eye movements and imposed movements because extra-retinal information contributed significantly to saccadic localization.

Extra-retinal information about the time of saccadic occurrence is used to suppress sensations of visual apparent motion, which would otherwise occur with saccadic image motion on the retina. This saccadic motion suppression is similar to re-interpretation of sensory inputs following voluntary movement in other modalities. The extent to which subjects can learn to ignore extra-retinal information in making visual judgements during saccades is not yet known.

Other than the ability to compute apparent motion between related images (based on correlations between elementary local features), subjects have no special memory for images that are related by simple translation. To coordinate images produced by successive fixations, the visual/cognitive system needs spatial information about the direction of gaze. This directional information cannot easily be extracted from the image sequence but is normally provided by the oculomotor system in conjunction with the head and body. Possibly even this oculomotor directional information could be replaced by equivalent directional information acquired from other modalities.

While saccades are usually remarkably efficient, it was possible to create a search task in which performance was substantially improved by eliminating saccades and presenting stimuli at a rate five times faster than saccades (25 new search arrays per second). Sudden onsets of stimulation such as might be caused by saccades were shown not to be necessary to initiate information-processing episodes; gradual ramp onsets served equally well.

Most of these results should not have been surprising – hindsight is easier than insight – from the point of view of processing efficiency. Both the visual sensory system and the oculomotor system have evolved extremely specialized and extraordinarily sensitive processing capacities near their respective receptors and effectors. In the brain, however, motor signals concerned with eye movements and visual signals, the result of post-retinal image processing, apparently interact only at high levels where the visual signal, at least, is far removed from its sensory origin. Because visuo-motor interactions occur at a high level, it suggests that they may be modifiable and substitutable. For example, when the extent of saccadic eye movements is optically modified, visuo-motor recalibration quickly occurs.

The hypothesis that emerged was that direct sensory control of vision by the oculomotor system is unnecessary. For example, in order to avoid confounding motion signals produced by eye movements with real object motion, it is not necessary to desensitize the retina during saccades. It is sufficient to process all such motion signals equally, and then to disregard saccadic motion outputs at what might be regarded as an 'interpretive' level. On the other hand, to avoid noticing saccadic motion smear, the visual system has evolved to ignore smear-followed-by-clear signals independently of how they are generated. Again, there is no visuo-motor interaction here, merely an effective adaptation of the visual system to a mode of seeing.

Acknowledgements

The preparation of this chapter was supported by The Air Force Office of Scientific Research, Life Sciences Directorate, Visual Information Processing Program Grant 88-0364 and by the Office of Naval Research, Cognitive and Neural Sciences Division, Grant N00014-88-K-0569. The author wishes to express his appreciation for the assistance provided by the late Roseanne G. Speelman in the experiments reported herein.

References

Adelson, E.H. and Bergen, J. (1985) Spatiotemporal energy models for the perception of motion. J. Opt. Soc. Am. A, 2, 284–299.

Averbach, E. and Sperling, G. (1961) Short term storage of information in vision. In C. Cherry (Ed.), Information Theory. Butterworths, Washington DC, pp 196–211

Ballard, D.H. (1987) Eye movements and spatial cognition. Technical Report 218, University of Rochester Computer Science, 50 pp

Bischof, N. and Kramer, E. (1968) Untersuchungen und Ueberlegungen zur Richtungswahrnehmung bei Willkuerlichen sakadischen Augenbewegung. Physiol. Forsch 32, 185–218.

Bridgeman, B. and Mayer, M. (1983) Failure to integrate visual information from successive fixations. Bull. Psychonomic Soc. 21, 285–286.

Bridgeman, B., Hendry, D. and Stark, L. (1975) Failure to detect displacement of the visual world during saccadic eye movements. Vision Res. 15, 719–722.

Brooks, B.A. and Fuchs, A.F. (1975) Influence of stimulus parameters on visual sensitivity during saccadic eye movement. Vision Res. 15, 1389–1398.

Brooks, B.A. and Impelman, D.M. (1981) Suppressive effects of a peripheral grating displacement during saccadic eye movement and during fixation. Exp. Brain Res. 42, 489–492.

Brooks, B.A., Impelman, D.M. and Lum, J.T. (1980a) Influence of background luminance on visual sensitivity during saccadic eye movements. Exptl. Brain Res. 40, 322–329.

Brooks, B.A., Yates, J.T. and Coleman, R.D. (1980b) Perception of images moving at saccadic velocities during saccades and during fixation. Exp. Brain Res. 40, 71–78

Bruno, N. and Cutting, J.E. (1988) Minimodularity and the perception of layout. J. Exp. Psychol. Hum. Percept. Performance, 117, 161–170.

Burr, D. (1980) Motion smear. Nature 284, 164–165.

Campbell, F.W. and Wurtz, R.H. (1978) Saccadic omission: why we do not see a grey-out during a saccadic eye movement. Vision Res. 18, 1297–1303

Clocksin, W.F. (1980) Perception of surface slant and edge labels from optical flow: a computational approach. Perception 9, 253–269

Collewijn, H., Steinman, R.M. and van der Steen, H. (1985) The performance of the smooth pursuit eye movement system during passive and self generated stimulus motion. J. Physiol. 366, 19P.

Corfield, R., Frosdick, J.P. and Campbell, F.W. (1978) Grey-out elimination: the roles of spatial waveform, frequency and phase. Vision Res. 18, 1305–1311.

Davidson, M.L., Fox, M.J. and Dick, A.O. (1973) Effect of eye movements on backward masking and perceived location. Percept. Psychophys. 14, 110–116.

Ditchburn, R.N. and Ginsborg, B.L. (1952) Vision with a stabilized retinal image. Nature 170, 36–37.

Dodge, R. and Fox, J.C. (1928) Optic nystagmus. I. Technical introduction with observations in a case with central scotoma in the right eye and external rectus palsy in the left eye. Arch. Neurol. Psychiatry 20, 812–823.

Dosher, B.A., Sperling, G. and Wurst, S.A. (1986) Tradeoffs between stereopsis and proximity luminance covariance. Vision Res. 26, 973–990

Dubois, M.F.W. and Collewijn, H. (1979) Optokinetic reactions in man elicited by localized retinal motion stimuli. Vision Res. 19, 1105–1115

Efron, R. (1970a) The relationship between the duration of a stimulus and the duration of a perception. Neurophysiologica 8, 37–55

Efron, R. (1970b) The minimum duration of perception. Neurophysiologica 8, 56–63

Eriksen, C.W. and Collins, J.F. (1967) Some temporal characteristics of visual pattern perception. J. Exp. Psychol. 74, 476–484

Feldman, J.A. (1985) Four frames suffice: a provisional model of vision and space. Behav. Brain Sci. 8, 265–289.

Fisher, D.L. (1982) Limited-channel models of automatic detection: capacity and scanning in visual search. Psychol. Rev. 89, 662–692.

Gerrits, H.J.M. and Vendrik, A.J.H. (1974) The influence of stimulus movements on perception in parafoveal stabilized vision. Vision Res. 14, 175–180.

Haegerstrom-Portnoy, G. and Brown, B. (1979) Contrast effects on smooth-pursuit eye movement velocity. Vision Res. 19, 169–174

Hansen, P.M. (1979) Spatial localization during pursuit eye movements. Vision Res. 19, 1213–1221.

Hansen, R.M. and Skavenski, A.A. (1977) Accuracy of eye position information for motor control. Vision Res. 17, 919–926

Hansen, R.M. and Skavenski, A.A. (1985) Accuracy of spatial localizations near the time of saccadic eye movements. Vision Res. 25, 1077–1082.

Heeger, D.J. (1987) A model for the extraction of image flow. J. Opt. Soc. Am. A, 4, 1455–1471

Hochberg, J. (1968). In the mind's eye. In R.N. Haber (Ed.), Contemporary Theory and Research in Visual Perception. Appleton-Century-Crofts, New York.

Hoffman, D.D. (1982) Inferring local surface orientation from motion fields. J. Opt. Soc. Am. 72, 888–892.

Hogben, J.H. and Di Lollo, V. (1974) Perceptual integration of brief visual stimuli. Vision Res. 14, 1059–1069.

Irwin, D.E., Yantis, S. and Jonides, J. (1983) Evidence against visual integration across saccadic eye movements. Percept. Psychophys. 34, 49–57.

Irwin, D.E., Brown, J.S. and Jun-Shi Sun. (1988) Visual masking and visual integration across saccadic eye movements. J. Exp. Psychol. Gen. 117, 276–287.

Javel, L.E. (1878) Essai sur la physiologie de la lecture. Ann. d'Oculistique 82, 242–253.

Jonides, J., Irwin, D.E. and Yantis, S. (1982) Integrating visual information for successive fixations. Science 215, 192–194.

Jonides, J., Irwin, D.E. and Yantis, S. (1983) Failure to integrate visual information for successive fixations. Science 222, 188.

Keller, E.L. and Zee, D.S. (Eds.) (1986) Adaptive Process in Visual and Oculomotor Systems. Pergamon Press, Oxford

Kelly, D.H. (1979) Motion and vision. II. Stabilized spatio-temporal threshold surface. J. Opt. Soc. Am. 69, 1340–1349.

Khurana, B. and Kowler, E. (1987) Shared attentional control of smooth eye movement and perception. Vision Res. 27, 1603–1618

Koenderink, J.J. and van Doorn, A.J. (1986) Depth and shape from differential perspective in the presence of bending deformations. J. Opt. Soc. Am. A, 3, 242–249.

Kowler, E. and Sperling, G. (1980) Transient stimulation does not aid visual search: implications for the role of saccades. Percept. Psychophys. 27, 1–10.

Kowler, E. and Sperling, G. (1983) Abrupt onsets do not aid visual search. Percept. Psychophys. 34, 307–313.

Kowler, E. and Steinman, R.M. (1980) Small saccades serve no useful purpose: reply to a letter by R.W. Ditchburn. Vision Res. 20, 273–276.

Kowler, E., van der Steen, J., Tamminga, E.P. and Collewijn, H. (1984) Voluntary selection of the target for smooth eye movement in the presence of superimposed, full field stationary and moving stimuli. Vision Res. 24, 1789–1798.

Krauskopf, J. (1957) Effet of retinal image motion on contrast threshold for maintained vision. J. Opt. Soc. Am. 47, 740–744

Krauskopf, J. (1960) Effect of target oscillation on contrast resolution. J. Opt. Soc. Am. 50, 1306.

Krauskopf, J. (1962) Light distribution in human retinal images. J. Opt. Soc. Am. 52, 1046–1050

Lennie, P. and Sidwell, A. (1978) Saccadic eye movements and visual stability. Nature 275, 766–768.

Loftus, G.R. (1972) Eye fixations and recognition memory. Cognitive Psychol. 3, 525–551.

Loftus, G.R. (1981) Tachistoscopic simulations of eye fixations on pictures. J. Exp. Psychol. Hum. Learn. Mem. 7, 369–376.

Longuet-Higgins, H.C. and Prazdny, K. (1980) The interpretation of a moving retinal image. Proc. R. Soc. Lond. Ser. B, 208, 385–397.

Mack, A (1970) An investigation of the relationship between eye and retinal image movement in the perception of motion. Percept Psychophys. 8, 291–298

Mack, A. (1986) Perceptual aspects of motion in the frontal plane In. K. Boff, L. Kaufman and J. Thomas (Eds.), Handbook of Perception and Performance, Vol. I, Wiley, New York, Ch. 17. pp. 1–38

MacKay, D M (1970a) Elevation of visual threshold by displacement of retinal image Nature 225, 90–92.

MacKay, D M (1970b) Mislocation of test flashes during saccadic image displacements. Nature 227, 731–733.

MacKay, D M (1973) Visual stability and voluntary eye movements. In. R Jung (Ed.), Handbook of Sensory Physiology, Vol 7 Springer Verlag, Berlin, pp. 307–331

Maloney, L.T. (1988) Spatially irregular sampling in combination with rigid movements of the sampling array. Invest Ophthalmol. Vis. Sci. 1988, 29, 58.

Maloney, L.T (1990) The consequences of discrete retinal sampling for vision In M S Landy and A J Movshon (Eds.), Computational Models of Visual Processing. Cambridge, MIT Press, Cambridge, MA

Mateeff, S. (1978) Saccadic eye movements and localization of visual stimuli Percept. Psychophys 24, 215–224

Matin, L. (1986) Visual localization and eye movements In K Boff, L. Kaufman and J Thomas (Eds.), Handbook of Perception and Performance, Vol 1, Wiley, New York, Ch 20, pp. 1–45

McConkie, G W and Rayner, K. (1976) Identifying the span of the effective stimulus in reading. literature review and theories of reading. In. H Singer and R B Ruddell (Eds.), Theoretical Models and Processes of Reading International Reading Association, Newark, DE.

McConkie, G W and Zola, D (1979) Is visual information integrated across successive fixations in reading? Percept Psychophys 25, 221–224

McConkie, G W Zola, D Blanchard H E. and Wolverton, G S (1982) Perceiving words during reading lack of facilitation from prior peripheral exposure Percept Psychophys 32, 271–281

Meyer, C H Lasker, A G and Robinson, D A (1985) The upper limit of human smooth pursuit velocity Vision Res 25 561–563

Murphy, B (1978) Pattern threshold for moving and stationary gratings during smooth eye movements. Vision Res 18, 521–530.

Nachmias, J. (1961) Determiners of the drift of the eye during monocular fixation J Opt Soc Am 51, 761–766

Neisser, U (1963) Decision time without reaction time experiments in visual scanning. Am J. Psychol 76, 376–385

Neisser, U (1964) Visual search Sci Am 210.

Neisser, U., Novick, R and Lazar, R (1963) Searching for ten targets simultaneously Percept Motor Skills 17, 955–961

O'Regan, J K. (1984) Retinal versus extraretinal influences in flash localization during saccadic eye movements in the presence of a visible background Percept Psychophys 36, 1–14

O'Regan, J K. and Levy-Schoen, A (1983) Integrating information from successive fixations: does trans-saccadic fusion exist? Vision Res 23, 765–768.

Palmer, J and Ames, C.T (1989) Measuring the effect of multiple eye fixations on size and shape discrimination Invest Ophthalmol. Vis Sc. ARVO Suppl. 30, 159

Pola, J. (1972) The relation of visual direction to eye position during and following a voluntary saccade. Unpublished doctoral dissertation. Columbia University.

Rathff, F and Riggs, L.A. (1950) Involuntary motions of the eye during monocular fixation. J. Exp Psychol 40, 687–701.

Rayner, K. and Pollatsek, K. (1983). Is visual information integrated across saccades? Percept Psychophys. 34, 39–48.

Rayner, K., McConkie, G W. and Zola, D (1980) Integrating information across eye movements. Cognitive Psychol 12, 206–226.

Reeves, A. (1977) The identification and recall of rapidly displayed letters and digits Unpublished doctoral dissertation, City University of New York.

Reeves, A. and Sperling, G (1986) Attention gating in short-term visual memory. Psychol Rev. 93, 180–206.

Riggs, L.A., Ratliff, F., Cornsweet, J C. and Cornsweet, T N. (1953) The disappearance of steadily fixated visual test objects. J. Opt Soc. Am 43, 495–501.

Shiori, S and Cavanagh, P (1989) Saccadic suppression of low-level motion Vision Res 29, 915–928

Shavenitk, A A., Hansen, R., Steinman, R.M and Winterson, B J. (1979) Quality of human retinal stabilization during small natural and artificial body rotations in man Vision Res 19, 675–683

Sperling, G (1960) The information available in brief visual presentations Psychol Monogr. 74, No. 11 (Whole No 498)

Sperling, G (1963) A model for visual memory tasks Hum Factors 5, 19–31.

Sperling, G. (1966) Comparisons of real and apparent motion J Opt. Soc. Am. 56, 1442.

Sperling, G. (1967) Successive approximations to a model for short-term memory Acta Psychol 27, 285–292.

Sperling, G (1970) Critical duration, supersummation, and the narrow domain of strength-duration experiments Behav. Brain Sc. 2, 279–282.

Sperling, G and Dosher, B A. (1986) Strategy and optimization in human information processing In K. Boff, L. Kaufman and J Thomas (Eds.), Handbook of Perception and Performance, Vol 1, Wiley, New York, Ch. 2, pp 1–65.

Sperling, G and Melchner, M J. (1978) The attention operating characteristic: some examples from visual search Science 202, 315–318

Sperling, G and Reeves, A. (1980) Measuring the reaction time of a shift of visual attention In R. Nickerson (Ed.), Attention and Performance, Vol VIII. Erlbaum, Hillsdale, NJ, pp. 347–360.

Sperling, G and Speelman, R G. (1964) Spatial localization during eye movements Am. Psychol. 19, 526–527.

Sperling, G. and Speelman, R.G. (1963) Visual spatial localization during object motion, apparent object motion, and image motion produced by eye movements. J. Opt. Soc. Am, 55, 1576

Sperling, G. and Weichselgartner, E. (1989) Movement dynamics of spatial attention Mathematical Studies in Perception and Cognition, 89–14. Department of Psychology, New York University.

Sperling, G., Budiansky, J., Spivak, J G. and Johnson, M C. (1971) Extremely rapid visual search: the maximum rate of scanning letters for the presence of a numeral. Science 174, 307–311.

Sperling, G., Landy, M S., Dosher, B A. and Perkins, M. (1989) Kinetic depth effect and identification of shape. J. Exp. Psychol. Hum. Percept. Performance 15, 826–840.

Sperling, G., Dosher, B A. and Landy, M.S. (1990) How to study the kinetic depth effect experimentally. J. Exp. Psychol. Hum. Percept. Performance 16, 445–450.

Stark, L., Kong, R., Schwartz, S., Hadzy, D. and Bridgeman, B. (1976) Saccadic suppression of image displacement. Vision Res 16, 1185–1187.

Steinman, R M. and Cushman, H. (1973) Binocular retinal image motion during active head rotation. Vision Res. 20, 415–429.

Steinman, R.M., Haddad, G M., Skavenski, A.A. and Wyman, D. (1973) Miniature eye movement Science 181, 810–819.

Steinman, R.M., Levinson, J Z., Collewijn, H. and Steen, J. van der (1985) Vision in the presence of known natural retinal image motion. J. Opt. Soc. Am. A, 2, 226–233.

Thurstone, L.L. (1927) A law of comparative judgment. Psychol. Rev. 34, 273–286.

van Santen, J.P.H. and Sperling, G. (1984) A temporal covariance model of motion perception. J. Opt. Soc. Am. A, 1, 451–473

van Santen, J.P.H. and Sperling, G. (1985) Elaborated Reichardt detectors. J. Opt. Soc. Am. A, 2, 300–321.

Vollmann, F.C. (1986) Human visual suppression. Vision Res. 26, 1401–1416.

Watson, A.B., Ahumada, A J. and Farrell, J.E. (1986) The window of visibility: a psychophysical theory of fidelity in time-sampled motion displays. J. Opt. Soc. Am. A, 3, 300–307.

Watson, A B. and Ahumada, A J. (1983) A look at motion in the frequency domain. NASA Tech. Memo. 84352.

Watson, A.B. and Nachmias, J. (1977) Patterns of temporal interaction in the detection of gratings. Vision Res. 17, 1143–1149.

Weichselgartner, E. (1984) Two processes in visual attention. Unpublished doctoral thesis Department of Psychology, New York University, pp 141.

Weichselgartner, E. and Sperling, G. (1985) Continuous measurement of visible persistence. J Exp. Psychol. Hum. Percept. Performance 11, 711–723.

Weichselgartner, E. and Sperling, G. (1987) Dynamics of automatic and controlled visual attention. Science 238, 778–780.

Westheimer, G., and Campbell, F.W. (1962) Light distribution in the image formed by the living human eye. J. Opt. Soc. Am. 52, 1040–1045.

Whipple, W.R. and Wallach, H. (1978) Direction specific motion thresholds for abnormal image shifts during saccadic eye movement. Percept. Psychophys. 24, 349–355.

Winterson, B J. and Steinman, R.M. (1978) The effect of luminance on human smooth pursuit of perifoveal and foveal targets. Vision Res. 18, 1165–1172.

Woodworth, R.S. (1906) Vision and localization during eye movements. Psychol. Bull. 3, 68–70.

Woodworth, R.S (1938) Experimental Psychology. Henry Holt and Co., New York.

Yarbus, A L. (1957) The perception of an image fixed with respect to the retina. Biofiz. (USSR) 2, 683–712. (Translated by J S. Bradley.)

Yellott, J I., Jr. (1982) Consequences of spatially irregular sampling for reconstruction of photon noisy images. Invest. Ophthalmol. Vis. Sci. ARVO Suppl. 22, 133.

Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 9631–9635, December 1989
Psychology

Texture interactions determine perceived contrast

(spatial vision/perception/lateral inhibition)

CHARLES CHUBB, GEORGE SPERLING, AND JOSHUA A. SOLOMON

Human Information Processing Laboratory, Center for Neural Sciences and Department of Psychology, New York University, 6 Washington Place, New York, NY 10003

Contributed by George Sperling, August 24, 1989

ABSTRACT For a patch of random visual texture embedded in a surrounding background of similar texture, we demonstrate that the perceived contrast of the texture patch depends substantially on the contrast of the background. When the texture patch is surrounded by high-contrast texture, the bright points of the texture patch appear dimmer, and simultaneously, its dark points appear less dark than when it is surrounded by a uniform background. The induced reduction of apparent contrast is greatly diminished when (*i*) the texture patch and background are filtered into nonoverlapping spatial frequency bands or (*ii*) the texture patch and background are presented to different eyes. Our results are unanticipated by all current theories of lightness perception and point to a perceptual mechanism for contrast gain control occurring at an early cortical or precortical neural locus.

Simultaneous Brightness Contrast. The perceived lightness of a uniformly luminant disc viewed on a large uniform surrounding background depends not directly on the luminance of the disc, but rather on the ratio of disc luminance to background luminance (1–3). Even a spatially restricted background affects perceived lightness as illustrated by the illusion shown in Fig. 1 *a* and *b*. The discs in Fig. 1 *a* and *b* are equiluminant, nonetheless, the disc in *a* appears lighter than the disc in *b*. This phenomenon of *simultaneous contrast* is interpreted in terms of a ratio rule by noting that in *a* the ratio of the disc's luminance to background luminance is greater than 1; in *b*, the ratio is less than 1.

Lateral Inhibition. A natural way to explain simultaneous contrast is in terms of *lateral inhibition*. Many models based on lateral inhibition have proposed that, at some level of visual processing, neurons strongly stimulated by the high-intensity background of the disc in Fig. 1*b* suppress the less strongly stimulated neurons responding to the interior of the disc. In Fig. 1*a*, the corresponding neurons within the disc receive no such inhibition from the weakly stimulated neurons surrounding them. Consequently, the neurons located within the disc of Fig. 1*a* respond more vigorously than their counterparts in 1*b*.

Under the crudest lateral inhibition model, the lightness of a given point in the visual field would be suppressed in proportion to the intensity of each nearby point (1). But such a scheme would result, for example, in lower lightness values for points near the edge of the disc in Fig. 1*b* than for points in its interior. The fact that both discs in Fig. 1 *a* and *b* appear to be of uniform lightness across their full expanse suggests a more complex form of lateral inhibition (4). Regardless of their details, all models that invoke the principle of lateral inhibition rest on the assumption that the primary factor determining the perceived lightness of either disc in Fig. 1 *a* or *b* is the ratio, at the disc edge, of disc luminance to background luminance.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Induced Reduction of Apparent Contrast. We report here an apparent lightness phenomenon that is beyond the scope of all such models. The basic effect can be observed in a display analogous to Fig. 1 *a* and *b*, except that—instead of varying luminance between a disc and its background—we vary the *contrast* of a random visual texture. In Fig. 1*c*, the zero-luminance (black) background of Fig. 1*a* becomes a zero-contrast (mean-luminance) gray field; the high-luminance white field of Fig. 1*b* becomes a high-contrast texture field in Fig. 1*d*; and the two gray discs become discs of intermediate texture contrast (0.5).

It is an empirical fact that all observers perceive the texture disc of Fig. 1*c* as being somewhat higher in contrast than the texture disc in Fig. 1*d*, despite the fact that the two discs are identical. (We describe a stronger form of the illusion below.) The bright pixels in the texture disc of Fig. 1*c* appear brighter than their counterparts in *d*, and simultaneously the dark pixels in the disc of *c* appear darker than their counterparts in *d*.

For each of the discs in Fig. 1 *c* and *d*, the average difference in luminance at the border between the disc and its background is 0 (except for random fluctuations). In fact, every single pixel in Fig. 1 *c* and *d* has an expected luminance equal to the mean luminance. Therefore, except for random fluctuations, any two areas of Fig. 1 *c* and *d* have the same average luminance, and any consistent difference in appearance between the discs of Fig. 1 *c* and *d* cannot be accounted for by standard (luminance-based) lightness models.

EXPERIMENTS 1 AND 2: CONTRAST AND LIGHTNESS INDUCTION

Method. To compare Fig. 1 *c* and *d*, most observers shift their eyes back and forth between the two texture discs. To produce a stronger version of the texture-contrast illusion that does not involve eye movements, we use just Fig. 1*d* and modulate the contrast of the background texture sinusoidally in time between extreme contrasts of 0 and 1. In addition, we produce a new, independent realization of the random pattern instantiated by Fig. 1*d* 60 times per second. This produces 60-Hz texture flicker over the whole field, but it eliminates any figural cues and renders negligible any effects of eye movements on the spatiotemporal frequency content of the retinal stimulus. The slow contrast modulation of the background causes subjects to perceive the contrast of the texture disc to be modulating in antiphase. When background contrast is high, texture-disc contrast appears to be low, and vice versa.

We used two nulling experiments to measure the induced modulation of the apparent lightness of both the dark and bright pixels of the texture disc. In the first nulling experiment, subjects viewed the texture disc while the contrast of the surrounding background was being sinusoidally modulated (at 0.47 Hz) between 0 and 1. Simultaneously, the contrast of the center disc was modulated in phase with the background. The mean luminance of the texture disc was

Arch- 14

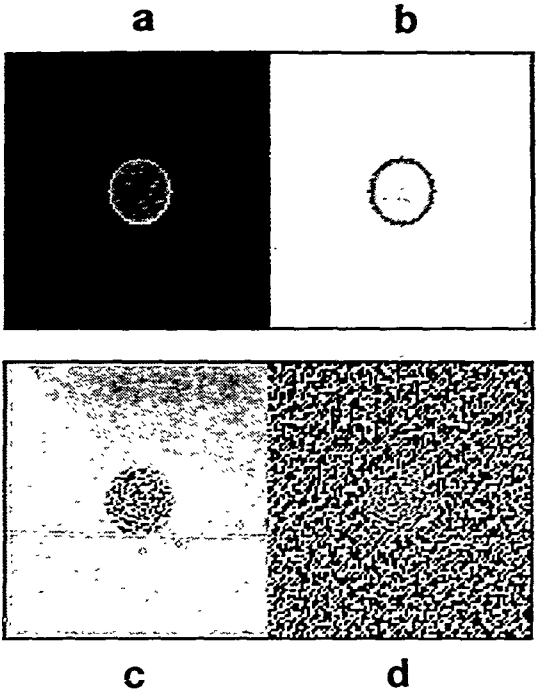


FIG. 1 Two factors influencing the lightness values assigned points in the visual field. (*a* and *b*) Classical lightness contrast. The lightness of a disc viewed on a background depends not only on the luminance of the disc but also on the ratio of disc luminance to background luminance (boundary contrast). The ratio of the luminance of disc to background is greater than 1 for *a* and less than 1 for *b*. Although discs in *a* and *b* have the same luminance, that in *a* appears lighter than that in *b*. (*c* and *d*) Induced contrast reduction. Like the mean-luminant discs in *a* and *b*, the texture discs in *c* and *d* are identical; each is of contrast ≈0.5. Because of the lower-contrast background, the disc in *c* appears to be of higher contrast than that in *d*.

kept constant in time. Subjects adjusted the modulation amplitude of the disc's contrast until disc contrast appeared constant in time.

The purpose of the second experiment was to determine whether or not there was a modulation of texture-disc overall lightness induced by modulating the contrast of the texture background. Accordingly, the contrast and the mean luminance of the texture disc were simultaneously modulated in phase with the background. The modulation amplitude of texture-disc contrast was fixed at the level (determined for each subject in the first experiment) at which the induced contrast modulation was nulled. Then, subjects adjusted the amplitude of texture-disc mean luminance modulation until the overall lightness of the disc appeared constant in time.

All displays were viewed binocularly from a chin rest at a distance of 1 m. At this distance, the texture disc was 1.35° in diameter centered in the 3.6° square background texture field.

Results. We tested texture discs with mean contrasts ranging from 0.2 to 0.5, and for all (*i*) the induced contrast modulation of the texture disc was substantial, while (*ii*) the induced overall lightness modulation was negligible. Thus, modulating the contrast of the texture background induces joint modulations of the apparent lightnesses of dark and bright pixels in the texture disc—joint modulations that are canceled by equal and opposite modulations of the luminances of dark and bright pixels in the disc.

The magnitude of this illusion is illustrated graphically in Fig. 2 for a mean texture-disc contrast of 0.4. The sinusoidal broken line gives the contrast of the background as a function of time. For a texture disc whose mean contrast (over time) is fixed at 0.4, subjects found it necessary (in the first nulling experiment) to modulate texture-disc contrast in accordance with the solid line of Fig. 2 in order to make texture-disc contrast appear constant in time. Thus, the texture disc appears to remain at a constant contrast (as shown by the flat broken line of Fig. 2) when its contrast is actually modulating in conformity with the solid line of Fig. 2. The amplitude of this nulling modulation (averaged for two subjects) is 45% of the texture disc's mean contrast. Similar data were obtained in other conditions.

EXPERIMENT 3: INTEROCULAR INDUCTION

Method. Is the induced modulation of texture-disc apparent contrast the result of an early or a late visual process? One way of investigating this question is to see whether or not the induction can occur across different eyes. Interocular induction implies that the neurons responsible for the induction must be at the level of the cortex or a higher visual center

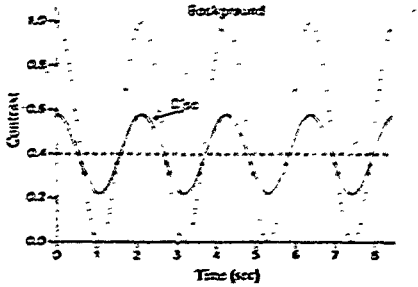


FIG. 2. Induced modulation of the apparent contrast of a texture disc. When a texture disc is viewed against a texture background whose contrast is sinusoidally modulated, the apparent contrast of the texture disc varies in antiphase to the background. In exp. 1, the background contrast varied between 0 and 1 at 0.47 Hz as shown by the sinusoidal broken line. For a texture disc whose mean contrast (over time) was fixed at 0.4, subjects found it necessary to modulate texture-disc contrast in accordance with the solid line to null the induced contrast modulation (i.e., to make contrast of the disc appear to follow the flat broken line). The amplitude of the nulling modulation was 45% of the texture disc's mean contrast.

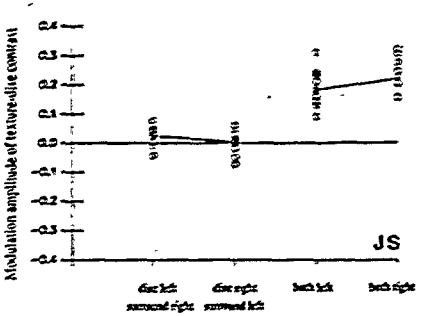


FIG. 3. Does the induced contrast modulation occur when texture disc and background are presented to opposite eyes? The subject modulates the contrast of a texture disc between two values around a mean of 0.4 so that it appears constant when the disc is viewed alternately against a high-contrast texture background and a mean-luminance background. Ordinate indicates the difference (resulting from subject adjustments) in texture-disc contrast between high-contrast-background presentations and mean-luminance-background presentations. Abscissa indicates the eye to which the disc and background are presented. Data (10 trials per condition) are shown for one subject (JS). Lines are drawn between the means. Only same-eye presentations induced a reduction of apparent contrast; this indicates an early cortical or precortical site for contrast gain control.

Strictly monocular induction implies that the locus of the induction is an early cortical or precortical cell population. Accordingly, we performed a third experiment in which the inducing background was delivered to one eye and the test disc to the other eye. Again we used the method of adjustment. There were four kinds of trials: (*i*) both disc and background were presented to the right eye; (*ii*) both were presented to the left eye; (*iii*) the left eye saw only the disc and the right eye saw only the background, and (*iv*) the right eye saw the disc and the left eye saw the background. Whenever a region of one eye's retina was presented with texture, the corresponding region of the opposite retina was presented with uniform mean luminance.

To minimize binocular rivalry, we used the following presentation sequence: The texture disc (which was 1.1° in diameter) was flashed periodically. Each flash lasted 133 ms, and flashes were separated by 500-ms periods of uniform mean luminance. Two types of disc flashes were alternated: background-on flashes and background-off flashes. On background-on flashes the texture disc was surrounded by a (2.9°) square texture background of contrast 1. On background-off flashes, the texture disc was surrounded by a background of contrast 0 (i.e., a uniform mean-luminant field). For some δ, under the subject's control, the contrast of the texture disc was $0.4 + \delta$ on each background-on flash and $0.4 - \delta$ on each background-off flash. On each trial, the subject adjusted δ (which was randomly initialized) until the contrast of the texture disc on background-on flashes appeared equal to its contrast on background-off flashes.

Results. Virtually identical data were obtained for two subjects; the data for one subject are shown in Fig. 3. On the trials in which both texture disc and texture background were presented to the same eye (either both to the right eye or both to the left), subjects had to make the contrast of the texture disc 40% higher on the background-on presentations than on the background-off presentations to equalize the apparent contrast of the texture disc across alternating background-on and background-off presentations. However, when texture disc and texture background were presented to opposite eyes, no such compensating adjustment was required. We infer that the contrast of the texture background influences the apparent contrast of the texture disc only when disc and background are presented to the same eye. This finding restricts the physiological location of the mechanism underlying this induction to an early cortical or precortical neuron population (5, 6).

EXPERIMENT 4: INDUCTION BETWEEN SPATIAL FREQUENCY BANDS

Method. In a fourth experiment, we examined whether or not texture filtered into one spatial frequency band could influence the perceived contrast of texture in a different spatial frequency band; that is, whether contrast induction is narrowly or broadly tuned for spatial frequency. We spatially filtered the texture of the disc through an ideal, octave-wide, nonoriented filter. The background was filtered by one of three adjacent octave-wide filters. The middle background filter was identical to the texture-disc filter (the frequencies passed by this filter were between 5.8 and 11.6 cycles per degree at a viewing distance of 1 m). Examples of each of the three textures are shown in Fig. 4.

Results. The results for two subjects are shown in Fig. 5. For both subjects, the largest contrast modulation is induced when the background texture is the same as the disc texture. When the background texture is in an adjacent octave-wide band, either one octave above or one octave below the disc texture, the induction is much weaker for both subjects. These results show that the reduction in apparent contrast of a disc induced by a textured background is spatial-frequency-specific. Preliminary investigations into orientation specificity indicate that when an oriented background texture is not in the same orientation as the disc texture, its influence on the perceived contrast of the disc texture is diminished.

DISCUSSION

The results of the fourth experiment suggest that, at some level of visual processing, neurons tuned to roughly a single octave (or less) in spatial frequency interact across space

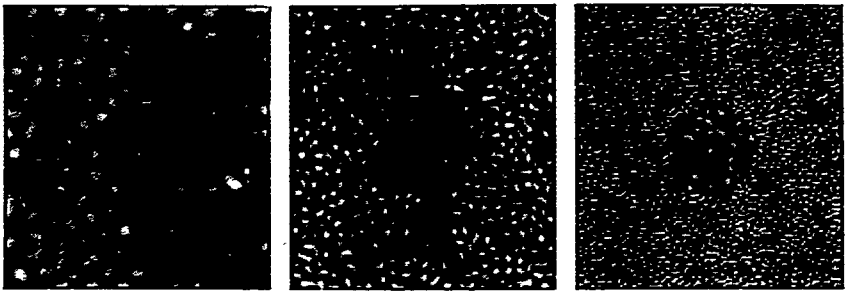


FIG. 4. Can texture of one spatial frequency affect the apparent contrast of texture of a different spatial frequency? Freeze frames of stimuli in which background and disc were filtered through ideal, octave-wide, spatial-frequency filters. (*a*) Background spatial frequencies are one octave below disc frequencies, (*b*) background and disc frequencies are the same, (*c*) background frequencies are one octave above disc frequencies.

with similarly tuned neurons. Taken together, our results support a model in which the output gain of such a band-selective neuron is normalized relative to the average response amplitude of nearby neurons with the same frequency tuning. Neurons differing in frequency tuning by more than an octave have much less influence on each other.

Several investigators have reported lateral inhibitory interactions between adjacent complex stimuli—for example, between textures of different spatial frequency (7), between lines differing in orientation (8, 9), and between different local velocities (10). The interactions have been small because these paradigms required the two stimuli to differ in their critical dimension, spatial frequency, orientation, or velocity. In a precursor of the present paradigm, Sagi and Hochstein (11) used a grating whose contrast was spatially modulated analogously to luminance in the Craik-O'Brien-Cornsweet illusion (12) to provide evidence for lateral texture-contrast inhibition. However, their display did not permit measurement of the effect.* Interneuron texture interactions have also been proposed on the basis of data obtained in searching for a target among distractor items (13). Prescient though such a theory may be, the data themselves admit other explanations and provide only indirect indications of texture interactions. Thus, the present experiments illustrate a kind of robust, spatial, feature-specific interaction that is (*i*) similar to gain control as observed in physiological experiments (14) and (*ii*) anticipated in the explanation of complex search tasks (13), but that has not, to our knowledge, been unambiguously observed before with simple textured stimuli in a psychophysical setting.

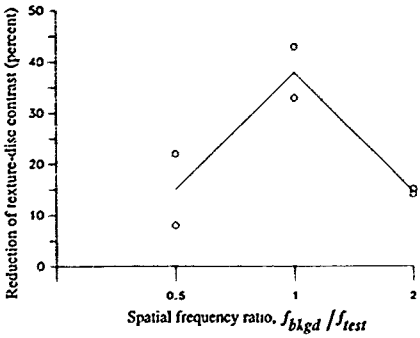


FIG 5. Induction of texture-disc apparent contrast is narrowly tuned for spatial frequency. A nulling procedure was used with the stimuli of Fig. 4. Ordinate indicates the difference in contrast between a texture-surrounded test disc (of contrast 0.4) and a texture disc matched in apparent contrast to the test disc, viewed against a uniform grey background. Abscissa indicates the spatial frequency of the background. Symbols indicate data for each of two subjects, each point is the average of the last 10 reversals of a staircase. Measurement error is approximately equal to symbol size. These data suggest that induced contrast reduction has approximately a one-octave spatial-frequency bandwidth

SUMMARY AND CONCLUSION

We have demonstrated the dependence of the perceived lightness of a point in space on lateral texture interactions in the visual display. The perceived contrast of a patch of texture is dramatically influenced by the contrast of surrounding texture. In particular, for spatial texture in a certain frequency band, the perceived contrast varies inversely with the contrast of surrounding texture in the same band. We showed that this lateral inhibitory effect is strictly monocular and that it is narrowly tuned for spatial frequency. The possible implications for perceptual theories are profound. On the one hand, it appears that the lightness of a point in space is a far more complex function of its environment than had hitherto been suspected—it will take a great deal of work to elaborate the precise spatiotemporal properties of the textural interactions sketched out here. On the other hand, if there are such specific lateral connections between spatial-frequency-tuned neurons and their similarly tuned neighbors, might there not be equally specific connections to normalize the responses of other classes of neurons? Is self-normalization a universal perceptual principle?

*Sagi and Hochstein also reported that a light bar of a grating adjacent to a zero-contrast area appeared lighter than other bars. It is possible to account for this effect in terms of simple luminance interactions; it does not strictly require texture interactions.

The authors are grateful to Barbara Dosher, Michael Landy, and Robert Shapley for their helpful comments. This work was supported

by Air Force Office of Scientific Research, Life Sciences Directorate, Vision Information Processing Program, Grant 88-0140.

1. Mach, E. (1866) *Sitzungsber. Math.-Naturwissens. Kl. Kaiser. Acad. Wiss.* 54, 131–144; Ratliff, F., trans. (1965) *Mach Bands: Quantitative Studies on Neural Networks in the Retina* (Holden-Day, San Francisco), pp. 272–284.
2. Hess, C. & Pretori, H. (1894) *Albrecht von Graefes Arch. Ophthalmol.* 40, 1–24.
3. Wallach, H. (1948) *J. Exp. Psychol.* 38, 310–348.
4. Grossberg, S. & Todorovic, D. (1988) *Percept. Psychophys.* 43, 241–277.
5. Hubel, D. H. & Wiesel, T. N. (1972) *J. Comp. Neurol.* 146, 421–450.
6. LeVay, S., Wiesel, T. N. & Hubel, D. H. (1980) *J. Comp. Neurol.* 191, 1–51.
7. MacKay, D. M. (1973) *Nature (London)* 245, 159–161.
8. Gibson, J. J. (1937) *J. Exp. Psychol.* 20, 553–569.
9. Blakemore, C., Carpenter, R. H. S. & Georgeson, M. A. (1970) *Nature (London)* 228, 37–39.
10. Walker, P. & Powell, D. J. (1974) *Nature (London)* 252, 732–733.
11. Sagi, D. & Hochstein, S. (1985) *Percept. Psychophys.* 37, 315–327.
12. O'Brien, V. J. (1958) *J. Opt. Soc. Am. A.* 48, 112–119.
13. Sagi, D. (1989) *Invest. Ophthalmol. Visual Sci.* 30, 161.
14. Shapley, R. M. & Victor, J. D. (1979) *Vision Res.* 19, 431–434.

Spatial Vision, Vol 4 Nos 2/3 pp 183-207 (1989)

Three stages and two systems of visual processing

GEORGE SPERLING
Human Information Processing Laboratory, Department of Psychology and Center for Neural Sciences, New York University, New York, NY 10003, USA

Received for publication 1 July 1989

Abstract—Three stages of visual processing determine how internal noise appears to an external observer: light adaptation, contrast gain control and a postsensory/decision stage. Dark noise occurs prior to adaptation, determines dark-adapted absolute thresholds and mimics stationary external noise. Sensory noise occurs after dark adaptation, determines contrast thresholds for sine gratings and similar stimuli, and mimics external noise that increases with mean luminance. Postsensory noise incorporates perceptual, decision and mnemonic processes. It occurs after contrast-gain control and mimics external noise that increases with stimulus contrast (i.e., *multiplicative* noise). Dark noise and sensory noise are frequency specific and primarily affect weak signals. Only postsensory noise significantly affects the discriminability of strong signals masked by stimulus noise; postsensory noise has constant power over a wide spatial frequency range in which sensory noise varies enormously.

Two parallel perceptual regimes jointly serve human object recognition and motion perception: a first-order linear (Fourier) regime that computes relations directly from stimulus *luminance*, and a second-order nonlinear (nonFourier) rectifying regime that uses the absolute value (or power) of stimulus *contrast*. When objects or movements are defined by high spatial frequencies (i.e., texture *carrier* frequencies whose wavelengths are small compared to the object size), the responses of high-frequency receptors are *demodulated* by *rectification* to facilitate discrimination at the higher processing levels. Rectification sacrifices the statistical efficiency (noise resistance) of the first-order regime for efficiency of neural connectivity and computation.

1. INTRODUCTION

Bandpass filtering refers to the processing of images or sounds so that they contain only a narrow range—typically one or two octaves—of component frequencies. In audition, bandpass filtering is used to create stimuli that stimulate only a small portion of the basilar membrane. By studying psychophysical responses to stimuli filtered in different bands, information processing mediated by each portion of the basilar membrane can be studied.

In vision, the aim of bandpass filtering is to create stimuli that stimulate only one or a small number of the visual channels that operate in parallel to process visual stimuli. Ideally, stimuli filtered in high frequency bands would stimulate only receptors (channels) with small receptive fields. Stimuli filtered in low frequency bands would stimulate only channels that have large receptive fields. (The term channel is used here to designate an information processing system characterized by receptors of a particular size.) As in audition, there is substantial interest not only in how stimuli that are confined to a single band are processed, but also in how information from stimuli in different bands is perceptually combined.

With the advent of affordable graphics processors, bandpass filtering has become an increasingly widespread stimulus manipulation in vision. Working with

bandpassed stimuli raises to the fore some important issues that are the subject of this article. With hindsight, we see, as usual, that some of these issues have been confronted before, but consideration of bandpassed stimuli offers important new insights. In other cases, new stimuli and procedures raise new questions and offer new opportunities. This paper coordinates data that have emerged from paradigms that utilize bandpass filtered stimuli together with a variety of other data in order to arrive at some general principles of sensory information processing.

2. VISUAL NOISE AT THREE STAGES OF PROCESSING

Consider first a study that was originally designed to determine whether image spatial frequencies or object spatial frequencies were critical for object discrimination. Parish and Sperling (1987a,b) filtered individual capital letters in five different spatial frequency bands (Fig. 1). They studied the role of three factors in the ability of subjects to identify these letters when they were embedded in noise: (1) the signal-to-noise ratio, (2) the object-relative spatial frequency band in which the letters-plus-noise were filtered and (3) the viewing distance (which determined retinal spatial frequency). They found that identification accuracy was independent of viewing distance over a range of more than 30:1. In this wide range, retinal spatial frequency did not matter in determining recognition accuracy; only object spatial frequency mattered. On the other hand, visual sensitivity to sine gratings at threshold varies enormously within the same range of retinal frequencies. In this section, we examine sine-grating detection and letter discrimination in order to define the various sources of noise that limit visual performance.

2.1. *Additive and multiplicative noise*

We consider two kinds of noise: additive noise and multiplicative noise. The term additive noise is used here to denote a stationary noise source that is independent of the signal and is added to the signal. Additive noise can be overcome by increasing

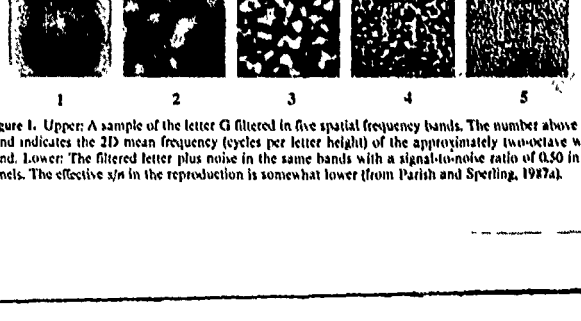


Figure 1. Upper: A sample of the letter G filtered in five spatial frequency bands. The number above the band indicates the 2D mean frequency (cycles per letter height) of the approximately two-octave wide band. Lower: The filtered letter plus noise in the same bands with a signal-to-noise ratio of 0.50 in all panels. The effective s/n in the reproduction is somewhat lower (from Parish and Sperling, 1987a).

signal strength until the effective signal-to-noise ratio is sufficient to support the desired level of performance

Multiplicative noise is proportional to the signal, that is, it multiplies the signal For example, in a binary (dark-grey/light-grey) image, reversing the contrast of (multiplying by -1) a randomly chosen 10 percent of the pixels would be a form of multiplicative noise Increasing the intensity or the contrast of the image would not alter its signal-to-noise ratio Multiplicative *noise is* equivalent to adding noise whose expected power is proportional to signal power Several authors have noted the distinction between additive and multiplicative noise (e g , Legge and Foley, 1980, Carlson and Klopfenstien, 1985, Legge *et al*, 1987, Pavel *et al*, 1987)

Loss of information that results from too-sparse sampling of the stimulus also can be regarded as a form of multiplicative noise (e g , Legge *et al*, 1987). To sample a signal means multiplying it by δ in the neighborhoods of the sampled points and by zero elsewhere (The constant δ is chosen so that the energy of the original signal is preserved in the sampled signal) For many applications, multiplying parts of a signal by zero (in sampling) is equivalent to multiplying parts of it by random numbers (*multiplicative noise*). In both cases, as the perturbation increases, *information* eventually is lost (noise), and the lost information cannot be recovered by increasing signal strength (*multiplicative* noise).

Multiplicative noise cannot be responsible for detection or discrimination thresholds that are reached by reducing the strength of a signal Because *multiplicative* noise declines proportionately with diminishing signal strength, weak signals are not worse off than strong signals. Because sufficiently low-luminance or sufficiently low-contrast visual signals are not visible, we infer that the internal noise that limits vision at low contrasts is better represented as additive rather than as multiplicative noise

There are many visual discrimination tasks in which increasing stimulus luminance or contrast does not improve performance. Consider six examples. In attempting to detect a spatial sine wave grating embedded in noise, with a fixed signal-to-noise ratio, the contrast of the display as a whole has no effect on performance once a critical contrast is reached (Pelli, 1981) *In detecting* spatial amplitude modulation of a one-dimensional spatial noise, once about eight times the contrast threshold for the noise is reached, further increases in overall contrast do not make the modulation more detectable (Jamar *et al*, 1982, Jamar and Koenderink, 1985). In Parish and Sperling's (1987b) letter-in-noise discrimination task, only the signal-to-noise ratio matters [1] In discriminating direction of motion, once a contrast of about 0.05 is reached, further increases in contrast do not improve performance (Nakayama and Silverman, 1985) Similarly, in audition, *typically the* signal-to-noise ratio (and not the absolute signal streng[ ]; determines performance For example, when a noisy radio broadcast is loud enough to be distinctly heard, making it louder does not make it more intelligible The visual analog, the independence of the discriminability of noisy, dynamic visual signals upon the stimulus contrast at which they were viewed was verified over a 4 1 range of contrasts by Pavel *et al.* (1987). In such cases, human performance appears to be characterized by multiplicative noise.

From a theoretical point of view, it is important to note that systems, which appear upon external examination to have identical multiplicative noise, may have vastly different internal mechanisms for generating their behavior. Viewed externally, the internal operation of *multiplying* the noise by a factor k before adding it to the signal

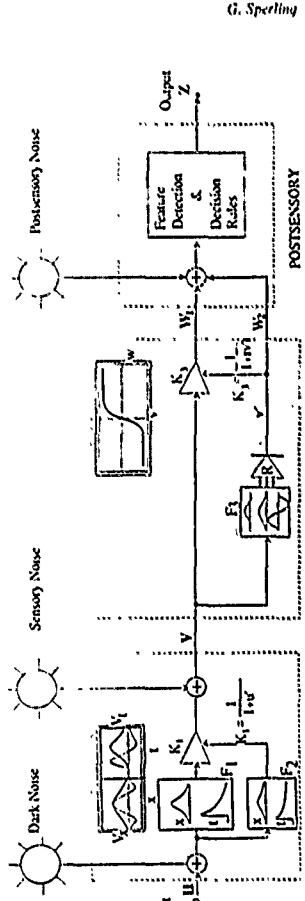


Figure 2. Three stages of visual processing. Sans indicate stationary noise sources. + indicates summation components, boxes indicate linear filters, triangles indicate amplifiers (gain control) and R indicates an array of rectifiers. Double boxes show input/output relations. *Adaptation.* The visual input is w the light-adapted output is v. The center/surround organization is produced by two pairs of separable spatial and temporal filters. F_1, F_2, respectively, whose impulse responses are indicated in the double boxes x and t. The surround F_2 controls the gain of amplifier K_1 to produce Weber law light adaptation (Sperling and Sondhi, 1968). The dark-adapted impulse responses are shown by the dark lines x_1 and t_1 vs. t. The light lines show the light-adapted impulse response relations. Dark noise adds directly to the input sensory noise adds after light-adaptation. *Contrast-gain control.* The graphs in box F_3 indicate various sigmoid and nonsinusoid spatial filters that operate on the light-adapted signal v. The full-wave rectified outputs, indicated by the triangle R, control the gain of the corresponding amplifiers K_2. Only one rectifier and amplifier are shown. A typical input–output function w vs. v is shown in the double-box insert. *Postsensory stages.* The first-order gain-controlled signal v_1, and the second-order gain-controlling signals u_1, are combined with each other and with noise. The postsensory stimulus, decision, mnemonic, and response processes are not detailed. The overall system output is z.

is equivalent to the internal operation of *dividing* the signal by k before adding it to the noise. Both result in the same internal signal-to-noise ratio. The equivalence of dividing signals by k and multiplying noise by k suggests gain control as a physiologically plausible internal mechanism to mimic multiplicative noise. The gain-control multiplies input signals by $1/k$ before a constant-power internal noise is added.

2.2. Three sources of visual noise

To understand how internal noise sources appear when viewed from the outside, it is useful to consider three stages of visual processing: light adaptation, contrast gain control, and postsensory processing (which includes perceptual, attentional, mnemonic, decision, and response processing). Figure 2 illustrates a flow chart for the computations carried out by these early stages. The particular mechanisms indicated in Fig. 2 for light adaptation and for contrast-gain control are based on physiologically plausible principles. They are vastly oversimplified and serve to illustrate the functional principles of the processes of light adaptation and gain control rather than the precise details (cf. Shapley and Enroth-Cugell, 1986). For example, the flow chart omits the division of signals into two distinct pathways that carry only positive and only negative signals (the on-center and the off-center neurons); parallel spatial frequency channels are not explicitly treated, there is no gain control for w_2, and so on.

Three stationary noise sources are illustrated in Fig. 2, each has constant expected power and an unchanging frequency spectrum. The three stages at which noise is added are (1) directly at the input, (2) after light adaptation and (3) after contrast-gain control.

2.3. Dark noise

In absolute darkness, the spontaneous activity of the visual receptors, rods and cones, is represented as dark noise (Barlow, 1956, 1957). Dark noise is prior to any processes responsible for light adaptation. To be reliably detected against a totally dark background, a signal must exceed not only the level of dark noise but also the combined level of all noise in the visual pathways. However, it would be expected that, through evolution, absolute threshold would be determined primarily by dark noise. That is, for receptors to serve most efficiently, their amplification gain would have increased (through evolution) up to the point where the receptor noise itself was the limiting factor.

2.4. Sensory noise

Sensory noise is the limiting noise in the detection of weak signals against uniform backgrounds. For example, by definition, a spatial sine wave grating with a contrast of 0.0001 has an absolute modulation that is proportional to its mean luminance. The brighter the illumination, the greater its absolute modulation. If there were no sensory noise, then increasing the absolute modulation of a spatial sinewave grating by increasing its mean luminance at constant contrast ultimately would increase its absolute modulation to the point of visibility (even with quantal noise in the stimulus). However, at high luminances, grating stimuli are visible very nearly in proportion to their contrast, not to their absolute modulation (Weber's Law). The essential fact of sensory noise is that, when viewed from outside the system at moderate to high intensities, apparent noise power increases with absolute modulation rather than

remaining constant. To model noise that apparently increases with the mean luminance (background intensity), the sensory noise source is placed after (central to) the gain control that modulates visual responsiveness as a function of intensity. Constant sensory noise, placed after the intensity gain-control mechanism, mimics an external additive noise that increases as a function of background intensity.

2.4.1. Sensory noise, Weber's law, quantal fluctuations. Weber's law asserts that the minimum detectable increment in intensity ΔS increases in direct proportion to background intensity S on which it is superimposed; at threshold: $\Delta S/S = k$, a constant. Assume that, at threshold, a constant signal-to-noise ratio is required at the detector itself. $s/n = [\text{signal amplitude}]/[\text{root-mean-square (RMS) noise amplitude}]$. Indeed, the effective signal-to-noise ratio of the stimulus at the detector is equivalent to the d' statistic of signal detection theory (Green and Swets, 1966). Internal noise after adaptation to the background is equivalent to external noise whose RMS power increased in proportion to the background intensity: either results in Weber's law behavior because, to maintain a constant signal-to-noise ratio, the threshold increment would have to increase in direct proportion to the mean background. Thus, sensory noise is assumed to be the source of Weber's law.

Most visual stimuli are produced by sources that can, for practical purposes, be approximated as quantal emitters. This means that, even with a nominally constant stimulus, the number of quanta collected by the retina in any given area varies from occasion to occasion and is characterized by a Poisson distribution. The *variance* of the Poisson distribution is equal to its mean; therefore, the RMS power of quantum noise increases in direct proportion to the *square root* of the luminance of visual stimuli. Because quantal noise increases with the stimulus amplitude, it usually is considered in conjunction with sensory noise.

The full analysis of quantal noise in the stimulus itself together with such factors as the blur of the visual optics and the spacing of retinal cone receptors is quite complex. For example, Banks *et al.* (1987) and Geisler (1989) applied such an analysis to contrast detection thresholds for sine wave stimuli of spatial frequencies from 5 to 40 c/deg, at mean luminances from 3.4 to 340 cd/m^2 (10.7 to 1068 trolands). Stimuli at each frequency consisted of seven sine cycles; i.e., the stimuli of different spatial frequencies were scaled replicas of each other. Once all the preneural factors cited above had been taken into account, at the observed thresholds, the stimulus s/n at the detector was constant (Banks *et al.*, 1988). The most parsimonious interpretation is that sensory noise is negligible compared to quantal stimulus noise for these stimuli. For sine wave gratings at lower spatial frequencies than 5 c/deg and for more intense stimuli at all spatial frequencies, sensory noise becomes quite significant relative to quantal noise. At very low levels of background luminance, dark noise becomes important (Geisler, 1989). Indeed, a model such as that of Fig. 2, together with threshold data obtained at different adaptation levels, offers a clear distinction between, and independent estimates of, residual sensory noise and dark noise.

2.5. Postsensory noise

In the suprathreshold experiments with added external noise discussed above, detection depended only on the signal-to-noise ratio s/n and not on the contrast at which these signals-plus-noise were viewed. In terms of a model, the dependence

of objective performance measures (such as direction-of-motion judgments, intelligibility scores, letter discriminations) on s/n and their independence of stimulus contrast is represented by a contrast-gain control that equates all signals that exceed a minimal contrast level. For example, the input/output function illustrated in the contrast-gain control box of Fig. 2 is shaped like a logistic function with an asymptotic output of -1 for large negative contrasts and an asymptotic output of $+1$ for large positive contrasts. A constant noise source that was located centrally (or added after) such a gain control would appear to an external observer to be equivalent to an external noise source that was directly proportional to contrast in those ranges of input where the gain-control was near its asymptotes.

From a functional point of view, all noise sources that are added after contrast-gain control will appear externally to be multiplicative noises, proportional to stimulus contrast. There are many such sources. Consider a two-alternative forced-choice intensity discrimination task. In successive intervals, an observer is presented with, for example, sounds of intensity 40 and 41 db and required to say which interval contained the louder sound. Generally, observers do better in a pure block of trials (only two sounds 40 and 41 db occur) than in a mixed block (e.g., trials with 40 and 41 db mixed in with trials containing 60 and 61 db, a 'roving' discrimination—Berliner and Durlach, 1973). In the pure block, the inability of the human observer to equal the performance of the ideal observer is attributed to a combination of (human) sensory and decision noise. In the mixed block, there is additional 'context' noise due to an attentional/mnemonic component. In an identification task, where observers must identify (name) each stimulus (e.g., 40, 41, 60, 61 db), their performance can be characterized as being further degraded by mnemonic noise.

The relative levels of performance in any two complex detection, discrimination, or identification tasks will be determined by a combination of shared noise sources and task-specific noise sources (e.g., MacMillan, 1987). All these postsensory noise sources are grouped together under the heading of postsensory noise, representing perceptual, contextual, decision, attentional, mnemonic and response processes that, according to the task, add noise after contrast-gain control.

To recapitulate. In vision, at threshold, sensitivity is governed by the intrinsic additive noise of the visual system (Pelli, 1981). Above threshold, matters apparently are quite different. "The notion of the observer's equivalent noise, which has been so useful in understanding detection, is found not to be relevant at suprathreshold contrasts." (Pelli, 1981, p. 121.) However, to formulate coherent theories of performance, we need merely to enlarge the concept of equivalent noise to include noise sources that, to an external observer, appear to vary with adaptation (because they are located after adaptation gain control) and noise sources that appear to vary with stimulus contrast (because they are located after contrast-gain control).

2.6 *The efficiency of detection*

The efficiency *eff* of discrimination is the ratio of s_i^2/n_i^2 required by an ideal observer to the s_h^2/n_h^2 required by a human observer at the same criterion level c of performance. $eff = (s_i/n_i)_c^2/(s_h/n_h)_c^2$. Alternatively, efficiency can be expressed as the squared ratio of human over ideal d' when confronted with the same stimuli. $eff = (d'_h/d'_i)^2$. Efficiency represents an estimate of how much less information than the human the ideal observer needs in order to match the human's performance. For example, in a visual display of n independent, equivalently informative pixels, *eff* is the fraction of the n pixels that the ideal observer needs to observe in order to match human performance.

Experimental determinations of efficiency establish an upper bound on the power of the human internal noise sources. Parish and Sperling (1987a) determined the efficiency of human discrimination in identifying visual letters masked by noise. When both the letters and noise were passed through a filter centered at 1.05 cycles per letter height, efficiency exceeded 0.40. Furthermore this high efficiency was observed over a 30:1 range of viewing distances. At the different viewing distances, these stimuli are transduced by visual channels characterized by vastly different retinal spatial frequencies. The constant high efficiency in a range where sensory noise varies enormously suggests that information loss in the visual pathway before the point of postsensory noise was negligible. In terms of noise sources, this means (1) that dark noise and sensory noise were negligible compared to stimulus noise and (2) the postsensory noise in the optimal band was of approximately the same power as the real stimulus noise (because the efficiency was near 0.5). Over the enormous range of spatial frequencies subserved by these channels, efficiency was determined primarily by postsensory noise.

3. LETTER DISCRIMINATION, NOISE AND THE SPATIAL MODULATION TRANSFER FUNCTION (MTF)

The MTF, also called the *contrast transfer function*, is the function that gives the contrast modulation of a sinewave grating at its threshold of detection as a function of its spatial frequency (Fig. 3). The question we address here is: How do the results of Parish and Sperling's letter discrimination experiments relate to what is already known about sine-wave detection? First, Parish and Sperling's (1987a) letter discrimination experiments involved an enormously greater range of low spatial frequencies than are typically used. In the range of retinal spatial frequencies for which data from both kinds of experiments are available, contrast threshold ranges from a minimum of 0.002 at 5–8 c/deg to a maximum of 0.07 at 37 c/deg (e.g., van Nes and Bouman, 1967, cf. Fig. 3). [The frequency of 37 c/deg is the mean retinal frequency of Parish and Sperling's highest frequency band 5d at its longest viewing distance, the most detectable retinal sine frequencies (5c/deg) are produced by Parish and Sperling's frequency bands 3d, 4c, and 5b (Fig. 3) at closer viewing distances.] In the letter-in-noise experiment, observed discrimination efficiency was independent of the mean retinal frequency (varying only with object spatial frequency) whereas, in the sine-wave grating detection experiment, threshold sensitivity for sinusoidal gratings varies from 0.07 to 0.002, a factor of 35, within the same frequency range (Fig. 3). Indeed, the combination of filter frequency with viewing distance in the letter-in-noise discrimination experiment produced retinal frequencies that varied over a range of more than 200:1 (Fig. 3), and discrimination was independent of retinal spatial frequency throughout this entire range.

Figure 3 illustrates the division of spatial frequencies into three regions:

(1) The top region which represents invisible sinusoidal gratings—their contrast is below detection threshold.

(2) A middle region, indicated in grey, in which detection is governed by quantal and sensory noise. In this region, increasing stimulus contrast improves performance.

(3) The lower region in which postsensory noise predominates. Here, noise is proportional to contrast so performance is independent of contrast. The numbers indicate the center frequency (projection on the x-axis) of various bandpass stimulus

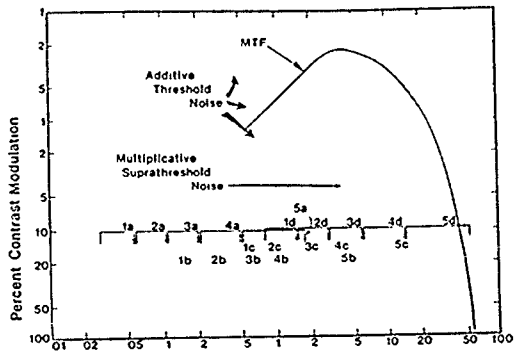


Figure 3 The contrast modulation transfer function (MTF) and the frequency ranges of the letter-in noise stimuli of the Parish and Sperling letter discriminaton experiments. The MTF gives the contrast detection threshold for sine gratings as a function of their retinal spatial frequency, it is based on data of van Nes and Bouman (1967). Stipling indicates the area, near threshold, where quantal noise and additive sensory noise predominate. Noise also predominates in the whole upper portion of the graph, where stimuli are invisible. Each open downward facing rectangle indicates the approximate retinal half-bandwidth of an object frequency band (1–5, Fig. 1) at one of the viewing distances (a = closest, d = furthest) used by Parish and Sperling (1987a). The horizontal placement of the corresponding number,letter symbol indicates the mean *retinal* frequency of the stimulus symbols (b, c) for intermediate viewing distances also are shown. The stimulus symbols are placed vertically at a contrast ≥ 10%, to indicate that for all larg contrasts (downward in the figure), performance is independent of contrast, i.e., it is controlled by multiplica e noise

conditions of the Parish–Sperling study, and the approximate contrast level (0.1, projection on the *y*-axis) at which performance becomes independent of stimulus contrast.

Previously, Jamar and Koenderink (1985) had noted an apparent independence of spatial-frequency in the detection of amplitude modulated noise gratings. They investigated a relatively small range of frequencies and did not determine the efficiency of detection. In letter detection, the enormous range of frequency invariance, and the extremely low level of decision noise (as demonstrated by comparison with ideal detectors) is truly astounding.

Detection thresholds for sine gratings vary enormously with retinal spatial frequency in precisely the same range of frequencies where the discrimination threshold for letters-in-noise is constant. The difference between the two experiments is readily interpreted in terms of the levels-of-noise model. The detection of low-contrast spatial gratings is limited by quantum noise in the stimulus and by sensory noise; the discrimination of letters-in-noise is limited by postsensory noise. Whereas letter-in-noise discrimination is unaffected by stimulus contrast over a wide range,

stimulus contrast is the dependent variable in the grating detection experiment. Indeed, the grating detection experiment can be viewed as indicating the effective power of quantal plus sensory noise as a function of spatial frequency. We say 'effective power' because there is no provision in the simple stage model for input amplification that may vary as a function of spatial frequency; input gain is incorporated into sensory noise.

3.1 Conclusions

Representing grating detection and letter-in-noise discrimination as noise limited processes yields the following conclusions: (1) Sine grating detection at low stimulus contrasts is limited by quantal noise (Banks *et al.*, 1988) and by sensory noise (Pelli, 1981) each of which varies little with stimulus contrast but varies greatly with retinal spatial frequency and with mean luminance. (2) Letter detection at stimulus contrasts greater than about 0.10 is limited by apparently multiplicative noise that is proportional to stimulus contrast but is independent of spatial frequency (for a 100-fold range of retinal spatial frequencies). (3) When letters are discriminated in external noise which deliberately is not negligible, the effective internal noise apparently varies multiplicatively with stimulus contrast. These empirical relationships follow from the stage model of Fig. 2; and they are illustrated in Fig. 3. (4) Sensory and postsensory noise are independent and vary differently with spatial frequency. For example, the channel that transduces 5c/deg has the lowest sensory noise, but it has the same decision noise as the channel that processes 37 c/deg, which has 35 times more sensory noise.

3.2 Analogous phenomena in psychoacoustics

A similar pattern of strong frequency dependence of threshold detection and frequency independence of high-intensity discrimination occurs in psychoacoustics. For example, absolute intensity-detection thresholds $\Delta I(f)$ for sinusoidal pressure waveforms vary enormously as a function frequency f. At high signal levels, detection thresholds for sinusoidal increments $\Delta I(f)/I$ hardly vary with frequency (Reisz, 1928; Robinson and Dadson, 1956; Jesteadt *et al.*, 1977; see Scharf and Buus, 1986, for a review). Detection limits at low input levels are quite different from discrimination limits at high input levels. The nature of these differences is dictated by requirements of having maximally sensitive receptors and of operating over an enormous dynamic range. Since these problems are shared by many modalities, we should not be surprised at functionally similar solutions.

3.3 Advantage of above-threshold gain that is independent of frequency

A visual object is characterized by relations between its component spatial frequencies. When the object is viewed from nearer or further, these relations do not change, they are merely transposed up or down the frequency axis. If the visual system had important gain differences between different spatial frequencies, then these differences would have to be incorporated into object descriptors in order to preserve object invariance with scale changes. Clearly, object description at a high level can be more economical when the low level description accurately represents the object's spatial frequency content. Constant gain across frequencies is the simplest way to begin a scale-invariant description.

An auditory object (e.g., a voice or a tune) is characterized by the relations between

the component auditory frequencies To a good approximation moving closer further from the source corresponds to an overall intensity change If cha intensity were to change the internal frequency relations the internal object desc would have to be intensity (distance) dependent an enormous complication threshold the internal representation of sounds (or visual stimuli) inevitably r the ear's (or the eye's) frequency sensitivity Above threshold it would be de for object descriptions to be intensity invariant Indeed the further above thre the less auditory and visual discriminations vary as a function of frequency

4 WHY YOU CAN'T SEE THE FOREST FOR THE TREES
THE ECONOMICS OF CONNECTIVITY

In audition frequencies above 20000 Hz are too high to be audible and amplifi will not make them audible In vision, apparently the opposite occurs For ex m Parish and Sperling's (1987a) letter discrimination task with two-octave-wide the higher the center frequency of the band, the more discriminable the lett there an upper visual object frequency at which the trend to improved discrimin reverses' What happens as visual stimuli are filtered in higher and higher frequency bands'

Consider first an ideal letter stimulus in which the letter has perfect zero-o edges (Fig 4b) When such an edge is bandpass filtered, for example, by a diff of Gaussians filter (Fig 4a) the results are alternate dark-light stripes centered edge (Fig 4c) When the frequency spectra of different filter bands are rela simple translation on a log frequency axis (i e , the frequency filters differ c scale) the basic shape of the bandpass filtered edge is independent of the frequency of the filter By suitably scaling the abscissa, we arrive at a cai representation like that of Fig 4(b) and 4(c) in which, as frequency band change the distance between edges changes, not the shape of the edge representation

Obviously the higher the filter's center frequency the narrower the edge repr tion But spatial bandpass filters operate on object - not retinal frequencies frequency of a band is made higher and higher, the viewer can preserve a c retinal frequency by approaching the letter closer and closer, ultimately, microscope Locally the edges filtered at different center frequencies f will p exactly the same retinal images when the viewing distance is proportional What changes with viewing distance is the retinal distance between opposit of a letter stroke In normal viewing, the thickness of a letter stroke may b arcmin With sufficient magnification, the width of the letter stroke ultimately to occupy degrees or even hundreds of degrees of visual angle For extreme frequency bands when a letter has been enlarged sufficiently to make its edges it is physically impossible to view the whole letter at one time This is the c problem of trying to read a newspaper under a high power microscope individual letters become unrecognizable

For the cases of letters printed with real ink on real paper, or illuminatec on real CRT screens, the scaling problem is similar to the case of ideal High spatial frequencies represent local texture information about ink drop paper fiber or about how a CRT screen is populated with spots of light-s phosphor local texture that obscures the larger landscape This is the pro being unable to see the forest for the trees (Sperling and Parish, 1985) The observation is inappropriate for the object being observed

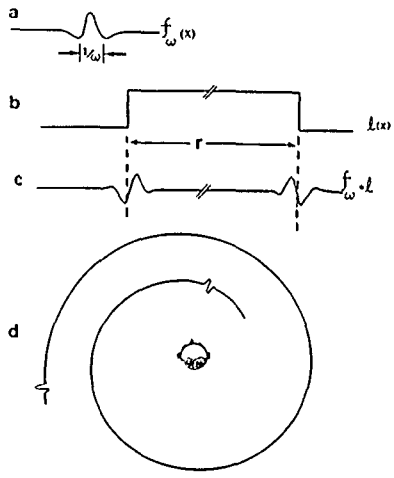


Figure 4 When is it impossible to see both the forest and the trees? The retinal image of a bandpass filtered boundary can remain independent of filter frequency f_ω by varying the viewing distance d For two boundaries that are physically separated by a distance ϕ the retinal distance r is ϕ/d Example (a) Impulse response of f_ω, a family of linear spatial filters that differ only in their scale ω (e g., ω is the center frequency of the passband). (b) A retinal illuminance distribution $l(x)$ representing the left and right boundary of an ideal white stripe (c) The retinal illuminance distribution of the filtered image $l * f_\omega$ with viewing distance d chosen so that $d = 1/\omega$ This choice of d normalizes (leaves unchanged) the images of the boundaries (neighborhoods of dashed lines in b and c). Only r varies with ω, not the boundary images (d) Ultimately for very large ω, r grossly exceeds 360 deg When it is necessary to view the boundary from extremely close in order to achieve visible detail, it will then be impossible to simultaneously resolve a boundary (see a tree) and to see both boundaries (see the forest)

4 1 *Economy of connection*

What is the appropriate scale of observation? This is dictated by the principle of *economy of connection* To compute relations, sensors must be connected to each other It is uneconomical for every sensor in a large field to be connected with and to compute its relations to every other sensor; typically sensors are connected only to immediate neighbors and to nearby neighbors A sensor and its similar neighbors form a kind of module The size of the visual receptive field viewed by a module is inversely related to the module's characteristic spatial frequency. In this arrangement, the optimal scale of observation is when the object is of the same order of size as the receptive field of the sensors so that the object can be entirely described within a module Indeed, the center frequency of the most efficient band for letter recognition

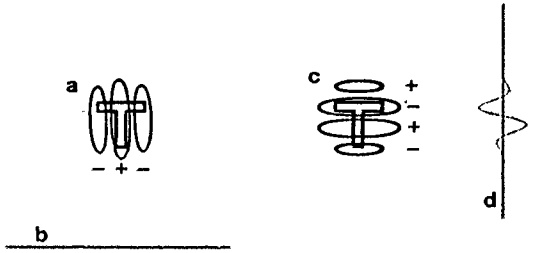


Figure 5. The letter T and receptive fields that have a center frequency of 1 cycle per letter height (in their higher frequency dimension). (a) The letter T centered in an even symmetric receptive field. The + and − signs indicate the sign of the field's response to spots of light in the indicated areas. (b) Horizontal cross section showing the sensitivity of the receptive field as a function of position. (c, d) The letter T within an odd-symmetric receptive field

is one cycle per object, i.e., the same order of size as the object. The size relation between letters and the spatial frequencies that were empirically found to be most efficient in identifying them is illustrated in Fig. 5. Note that several such spatial frequency filters, in different orientations and phases, would be required to discriminate between the 26 upper-case letters.

When, on the other hand, much smaller-sized sensors are used to describe a large object then, in a hierarchically organized system, it requires communication between modules, communication that occurs only at higher levels. Empirically, using high spatial frequencies to describe large objects results in a loss of perceptual efficiency. Below, we consider some reasons why communication *between* modules might entail a loss of information.

5. TWO PROCESSING SYSTEMS

The basic thesis of this section is that there are two processing systems: a Fourier system that uses phase information and makes local computations within a small local area (a module); and a nonFourier system that discards phase information and coordinates computations made in different modules. We approach these general issues by considering an analogy from radio communication.

5.1. *Demodulation*

5.1.1 High frequency carriers. In AM (amplitude modulated) radio communication, the amplitude of a high frequency *carrier* wave is modulated by the voice frequencies that are to be transmitted. Voice frequencies of up to about 10 000 Hz are transmitted as amplitude modulations of a 100 000 Hz carrier frequency. The process of extracting the low-frequency modulating signal from the high-frequency carrier frequency is *demodulation*

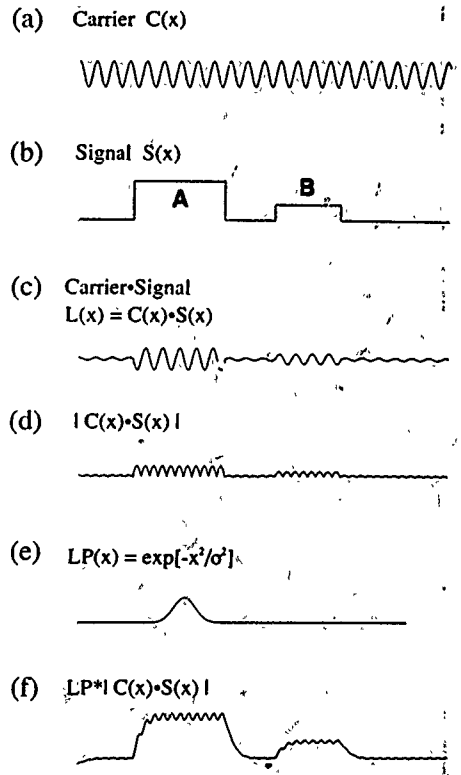


Figure 6. Carrier frequencies, amplitude modulation, and demodulation. (a) A carrier frequency $C(x) = \sin(2\pi f_c x)$. (b) A signal $S(x)$ that consists of an object A which has a large amount of the carrier f_c (one of its characteristic spatial texture frequencies), a second object B which has an intermediate amount, and the background which has a small amount. (c) A representation of the actual frequencies in the image, the image contrast distribution: $L(x) = S(x) \cdot C(x)$, an amplitude modulated carrier. (In visual scenes, the phase of the carrier is not preserved across objects.) (d) The rectified image. The absolute value of the image $|S(x)\ C(x)|$ is the simplest instantiation of fullwave rectification (e) A lowpass filter (Normal density function). (f) The result of lowpass filtering (d), $LP * |S(x)\ C(x)|$. (The * indicates convolution). The original signal $S(x)$ has been mostly recovered.

In visual object recognition, an analogous process of modulation occurs when an object A, whose overall shape is—by definition—characterized by frequencies around one cycle per object, is differentiated from its surround by higher frequencies. This would occur if the object had a surface texture that differed from the background texture. In that case, a spatial filter tuned to one of the dominant spatial frequencies in A, say f_a, would record a large response wherever A was present, and smaller responses elsewhere. Another object, *B*, might contain an intermediate amount of f_a (Fig. 6b) but a larger amount of other spatial frequencies. A texture functions, like a color, to characterize an object.

There is a close analogy between a characteristic texture frequency and an AM carrier frequency. The goal of demodulation is the same in both instances. In AM modulation, demodulation means estimating how much carrier signal (its amplitude) is present at each instant in time. In a texture-defined object, the problem for the visual system is estimating how much carrier signal is present at each point in space. The difference is that the phase of an AM carrier is consistent throughout the signal, the phase of texture-defined objects is not.

5.1.2 Neural mechanisms of fullwave and halfwave rectification. A simple form of demodulation involves fullwave rectification (taking the absolute value) of the signal (Fig. 6d). The modulated carrier is rectified and then lowpass filtered (Fig. 6e and f) to remove the carrier and higher frequencies; only the original modulating signal remains. In the visual system, after the initial receptors, positive and negative signals are carried in separate channels (e.g., on-center, off-center neurons). An alternative method of transmitting positive and negative quantities is to modulate the resting firing rate of a neuron up and down. The advantage of using separate positive and negative channels is that zero signal means zero impulses per second, and so the average firing rate is minimized.

When there are separate on- and off-channels, to preserve the sign of the signal at subsequent synapses, the target synapses for on- and off-neurons must operate in opposite directions (excitation or inhibition, see Fig. 7a). Fullwave rectification is accomplished when the target synapses of the on- and off-channels operate in the same direction (see Fig. 7b). Fullwave rectification means that the high-frequency sensors of the carrier frequency communicate information about their location and the magnitude (but not the sign) of their responses to the next higher level of the system. On the other hand, halfwave rectification (Fig. 7c) corresponds to independent analyses of the on- and off-channel signals, a process that has been proposed as a mechanism for locating luminance boundaries (Watt and Morgan, 1985).

Converting the output of high frequency detectors to lower frequencies (demodulation) is a critical component of object recognition because objects are defined most efficiently and most economically in the lowest feasible frequency range. The computational advantage of a hierarchical demodulatory scheme is that pattern recognition at the higher level can use a single computation that is independent of the scale or the contrast of the sensors that are transmitting information from lower levels. Because, in this context, demodulation involves going from higher to lower spatial frequencies, the pattern recognition algorithm can operate at the lowest frequency. Using the lowest possible frequencies is computationally efficient because of the economy of connection. A neuron and its immediate neighbors span the field of interest.

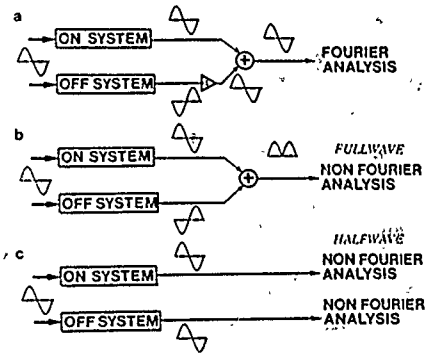


Figure 7. How linear transformations, fullwave rectification, and halfwave rectification can be accomplished in the visual system. *On-System* refers to neurons that have an on-center/off-surround receptive field organization (Kuffler, 1953) and that carry signals representing positive local contrasts relative to the surround. *Off-System* refers to off-center/on-surround neurons that transmit information about negative local contrasts. (a) When synapses from an On-System neuron onto a target neuron are excitatory and Off-System synapses are inhibitory (indicated by the inverting amplifier-Δ), the sign of input contrasts is preserved and first order (Fourier) analysis of the stimulus can occur. (b) Fullwave rectification occurs when both On- and Off-System synapses are the same (either excitatory or inhibitory); this results in *second order* signal analysis that is 'nonFourier'. (c) Positive halfwave rectification occurs when the On-System signals are analyzed independently; negative halfwave rectification refers to independent analysis of Off-System signals. Like fullwave rectification, halfwave rectification could be the essential nonlinearity in a second-order processing scheme.

In letter discrimination, the experimentally measured efficiency of discrimination was highest ($eff \approx 0.4$) at 1 cycle/object, the lowest usable band of spatial frequencies. Efficiency decreased to 0.1 at 10 cycles/object. Informational inefficiency is an unavoidable consequence of rectification because a computation that discards the sign of the input cannot be as efficient as one that takes sign into account. However, statistical inefficiency is a consequence of, not direct evidence for, demodulation or rectification. For direct evidence, we turn to other paradigms.

5.2. Direct evidence for two computational regimes in motion and texture-slant perception

5.2.1 The x, t cross-section of a motion stimulus. Perhaps the most convincing way to demonstate two computational systems is to embed two conflicting cues, one aimed at each system, in the same stimulus. The best examples occur in the domain of motion stimuli. The image of a moving stimulus is a three-dimensional (3D) function that gives luminance $I(x, y, t)$ as a function of x, y, t. To represent this 3D function on a printed page, we use x, t cross-sections that omit the y dimension as illustrated in

Figs. 8(a) and (b). Figure 8(a) shows a frame-by-frame representation of a rightward moving black bar; Fig. 8(b), shows the corresponding x,t cross-section. Superimposed on the bar's x,t cross-section in Fig. 8(b) is a sinewave. This sinewave is the x,t cross-section of a sinusoidal grating that is moving at the same velocity as the bar. This particular moving grating represents one of the largest Fourier components of the moving bar.

5.2.2. Motion stimuli that can be perceived in either of two directions. Figure 8(c) shows a space–time representation of a motion stimulus that has conflicting cues—a contrast-reversing bar, based on Anstis' (1970) reversed phi phenomenon. The bar steps sideways across a gray field, alternating its contrast between black (−1) and white (+1) on each step. The stimulus is Gaussian windowed in space–time so that only a few steps in the middle of the screen are maximally visible.

In the x,t cross-section, the bar moving to the right appears as a contrast-reversing diagonal slanting to the right. However, the Fourier sinewave components of the contrast-reversing bar are slanted down and to the left, indicating Fourier motion to the left. By rectifying the contrast-reversing bar, i.e., taking the absolute value of its contrast, the result is the stimulus of Fig. 8(b); and its Fourier sinewave components are slanted downward to the right, indicating rightward motion.

When such a contrast-reversing bar is viewed from near, it seems obviously to move to the right. However, when it is viewed in peripheral vision, or from a distance, or at very low contrast, it apparently moves to the left (Chubb and Sperling, 1988a, 1989b). This clearly indicates that observers make two different kinds of motion computations.

5.2.3. First-order motion perception. From an algorithmic point of view, a motion extraction system can be regarded as consisting of three consecutive types of computation: linear filtering, motion extraction, and decision. The first two, filtering and motion extraction, are carried out in parallel everywhere in visual space. A visual stimulus is processed first by linear filters in x,y,t that determine the range and amount of spatio-temporal frequencies that enter into the rest of the system, i.e., the linear filters determine the range of frequencies to which the motion system is sensitive. A second, inherently nonlinear stage, performs elementary local motion extraction by cross-correlation (Reichardt, 1957; van Santen and Sperling, 1984), spatio-temporal energy analysis (Adelson and Bergen, 1985), or some other relatively simple motion-extraction algorithm. At subsequent stages, the extracted motion from the various different motion detectors in each neighborhood and from various locations is compiled, and a decision is reached that is appropriate to the response demands.

For motion stimuli, Chubb and Sperling arrive at a functional discrimination between first-order (Fourier, direct) and second-order (nonFourier, rectified) processes.[2] The first-order regime is referred to as a Fourier process because it is well modeled by the type of computation described above. That is, when the stimulus contains Fourier motion components in the range of spatio-temporal frequencies to which humans are sensitive, the amount and direction of these Fourier components directly predicts the amount and direction of perceived motion.

5.2.4. Second-order motion perception. The contrast-reversing bar of Fig. 8(c) is an example of a class of stimuli for which the direction of perceived motion is opposite

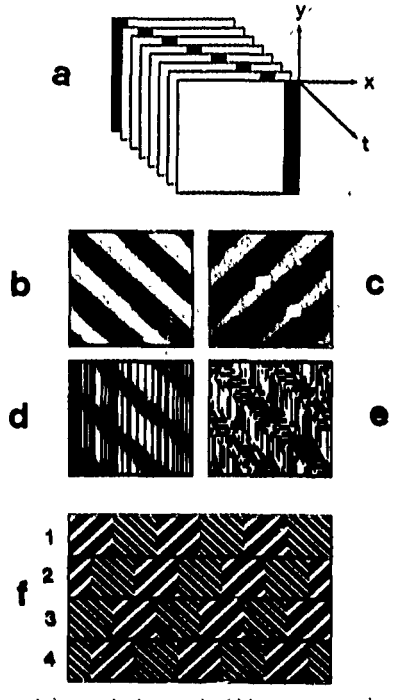


Figure 8. Stimuli for analyzing second-order processing. (a) An x,y,t representation of successive frames of a motion stimulus—a black bar moving rightward. (b) An x,t cross-section of (a). A sinewave grating, representing a dominant Fourier component, has been superimposed on the x,t cross-section. Note that the detection of direction of motion in x,t is equivalent to the detection of direction of slant in x,y. (c) An x,t cross-section of a windowed, contrast-reversing bar, a stimulus that appears to move leftward from afar (first order motion) and rightward from near (second-order motion). A sinewave grating, representing a dominant Fourier component, has been superimposed on the x,t cross-section to indicate the direction of Fourier movement. (d,e) x,t cross-sections of microbalanced stimuli whose motion is invisible to first-order motion detectors and whose slant in their x,y representation is invisible to first-order orientation detectors (e.g., Hubel-Wiesel cells) (f) A texture quilt. The four rows represent four successive frames of a dynamic stimulus. The initial extraction of either the low spatial-frequency texture oriented downward left or of the high frequency texture oriented downward right will enable a first-order motion algorithm to extract the overall leftward motion (overall slant downward to the left). Texture quilts remain microbalanced after any purely temporal transformation and require an initial texture extraction followed by textification to expose their motion in x,t (or orientation in x,y) to standard analysis.

to the direction of the Fourier motion components. Still other stimuli are perceived to move in the absence of any appropriate Fourier motion components. These 'non-Fourier' motion stimuli are detected by a second-order motion system. Second-order motion perception is now known to involve a stage of fullwave rectification (Fig. 7b) between the initial linear filtering and subsequent motion extraction. Second-order motion operates over larger retinal distances than does the first-order (Fourier) system. Additionally, consistent with the lower statistical efficiency of rectification, the second-order system has higher contrast thresholds than the Fourier system. Certain values of the parameters of viewing (such as small retinal size, peripheral retinal location, and low stimulus contrast) increase the relative strength of the first-order versus the second-order computation.

While the contrast-reversing bar is a simple demonstration stimulus, it does not enable one to discriminate between a fullwave and a halfwave second-order computation. Chubb and Sperling (1989b) demonstrate a sideways stepping, contrast-reversing grating, a stimulus which displays obvious second-order motion and in which halfwave rectification, alone or in combination with any reasonable temporal transformation, can be excluded. In displays that were designed to exclude fullwave rectification and admit only halfwave rectification or Fourier motion analysis, Sperling and Chubb (1987) did not find significant second-order motion. Thus, the predominant mechanism of second-order motion perception involves fullwave rectification. Fullwave rectification also is the dominant mechanism in second-order texture-slant processing of the x,y patterns that represented the x,t cross-sections of the motion stimuli in their motion experiments.[3]

In motion perception, there is a well-established distinction between short-range and long-range motion processes (e.g., Braddick, 1974; Pantle and Picciano, 1976; Westheimer and McKee, 1977; Victor and Conte, 1989b). The inadequacy of first-order motion processing has been amply documented by Ramachandran *et al.* (1973), Sperling (1976), Lelkens and Koenderink (1984), Pantle and Turano (1986), and Victor and Conte (1989a). The properties adduced for the long-range motion perception are generally those described for second-order motion above, plus a relative insensitivity to the eye of origin of successive stroboscopic stimuli. To these can be added the observations of Dosher *et al.* (1989) and Landy *et al.* (1987) that first-order motion supports the kinetic depth effect (KDE, Wallach & O'Connell, 1953) whereby 3D structure is perceived in 2D moving stimuli, whereas KDE induced by second-order motion stimuli is weak and of enormously lower resolution (e.g., Prazdney, 1987).

The computations of first-order motion are well embodied in the quite similar models of Watson and Ahumada (1983), van Santen and Sperling (1984), and Adelson and Bergen (1985). To supplement these theoretical proposals, van Santen & Sperling (1984, 1985) generated complex stimuli in which the direction and amount of perceived motion was opposite to intuition. Experimentally measurements of the perceived motion were quantatively predicted by their first-order, elaborated Reichardt model.

Chubb and Sperling (1988b, 1989a, b) provided a computational specification of a second-order motion system. They also provided methods for producing stimuli that can be proved mathematically to be directly aimed at one or the other system. One consequence is that it was easily shown that (retinal) short-range and long-range are inadequate system descriptions because there is a broad intermediate range in which both computations operate.

5.2.5. The equivalence of x,y spatial slant to x,t velocity. The velocity of a moving vertical line is the slope of its x,t cross-section (Fig. 8a, b). The slant of the line is the angle corresponding to the slope: slant = $\tan^{-1}$(slope). For any stimulus, velocity in x,t and slant in x,y are equivalent up to a simple monotonic transformation. The problem of left-vs.-right direction of motion discrimination for x,t motion stimuli involves formally identical computations to the problem of left-vs.-right slant discrimination for x,y spatial texture stimuli (van Santen and Sperling, 1985; Chubb and Sperling, 1987, 1988b).

To extract local texture slant in scenes, Knutsson and Granlund (1983) proposed a computation in which similarly slanted odd and even linear filters (line and edge detectors) added their squared (rectified) outputs. The core of their algorithm anticipates Adelson and Bergen's (1985) remarkably similar motion algorithm, which in turn, was shown (by van Santen and Sperling, 1985) to be equivalent to other, previously proposed motion mechanisms, including the Reichardt correlation detector (e.g., van Santen and Sperling, 1984) and an elaborated Watson and Ahumada (1983) detector. Thus, in confronting first-order motion stimuli, the theories of Fourier-based motion perception and Fourier-based texture-slant perception have followed a parallel evolutionary path to the same pinnacle of success; when confronted with second-order motion and texture stimuli, the theories arrive at the same abyss.

5.2.6. Drift-balanced and microbalanced second-order motion/texture stimuli. Strong evidence for two computational regimes is obtained in studies of slant detection in textured x,y patterns as well as in studies of direction discrimination in one-dimensional x,t motion perception. Figures 8(d) and (e) show demonstrations of stimuli that show obvious apparent motion (when presented as x,t motion stimuli) and obvious slant (orientation) when presented in x,y, as in the illustration.

The stimuli of Fig. 8(d) and (e) are driftbalanced: that is, they are exemplars of random stimuli in which the *expected* motion (or slant) is exactly equal for every pair of oppositely-directed component Fourier frequencies (Chubb and Sperling, 1988b). The overlaid sine gratings of Fig. 8(b) and (c) are an example of two oppositely-directed Fourier components—their slants in the x,t cross-sections are equal and opposite. The stimuli of Fig. 8(d) and (e) are not only driftbalanced, they are also microbalanced. This means, roughly, that every little area, whatever its shape, in these stimuli is driftbalanced. This means that the obvious slant in these x,y stimuli would be invisible to every linear Hubel–Wiesel cell (i.e., neurons with receptive fields such as illustrated in Fig. 5). The motion in the x,t versions of the stimuli is invisible to any standard (Fourier or Reichardt-equivalent) motion detector. Any oriented x,y filter or oriented x,t motion detector would have the same expected response for any microbalanced stimulus (in particular, zero, for the stimuli of Fig. 8d and e) as for their mirror-images. Rectification is required to make the x,t motion or x,y slant in these stimuli accessible to standard slant or motion analysis (Chubb and Sperling, 1988).

5.2.7. Texture quilts prove linear filtering precedes rectification. Figure 8(f) shows an example of a texture quilt (Chubb and Sperling, 1989a). The essential ingredient of a texture quilt is a moving patch of texture. The trick is that the spatial phase of the texture in the moving patch is uncorrelated from frame-to-frame.

The temporal version of texture quilt illustrated in Fig. 8(f) exhibits consistent

apparent motion; and the spatial version exhibits consistent slant. Although the adjacent patches in the quilt of Fig. 8(f) differ both in spatial frequency and in slant, preliminary experiments indicate that apparent motion is perceived in texture quilts with patches differing only in spatial frequency or only in slant. These results mean that the early texture grabbers are spatial-frequency selective and that at least some are orientation specific.

Chubb and Sperling (1990) prove that to make the overall motion in such a texture quilt accessible to motion analysis requires an initial stage of selective spatial filtering (texture grabbing) followed by rectification and standard motion (or texture) analysis. No purely temporal transformation, no matter how complex and nonlinear, can make this motion (or texture) accessible to first-order analysis. In terms of a perceptual computation, this means that the texture with which each patch of the texture quilt is filled must first be extracted and rectified before the information can be used in a larger-scale slant-extraction or motion-extraction computation to reveal the overall pattern.

Since the work of Schade (1952) and DeLange (1954), first-order Fourier-based computations have been the cornerstone of psychophysical analysis [4] The examples of Fig. 8(c, d, e) show the limitations of first-order linear analysis and the necessity of postulating second-order computations. Texture quilts (Fig. 8f) provide a fine tool for studying the spatial properties of the second-order processes for perceiving motion and texture-slant.

5.3 *Direct evidence for two computational regimes in distance judgments*

As with motion perception and texture-slant perception, distance estimation experiments also yield evidence for two perceptual processing systems, one Fourier and one rectifying system. In a three-bar distance estimation experiment, an observer must judge whether a central line is equally spaced between two flanking bars (bisection). In a two bar task, the observer directly judges the distance between two widely separated bars.

5.3.1 Direct estimates of distance. On a grey background, for widely separated bars, it matters not whether any of the bars is black and the other white, or whether both are white or both are black (Burbeck, 1987). Indeed, when 'bars' are defined by patches of high frequency gratings so that the bars themselves do not differ in average luminance from their surround, distance judgments are as accurate as with solid bars.

That observers accurately judge the distance between widely separated grating patches virtually guarantees a demodulatory process by means of which the stimulus grating patch is represented internally as a solid patch. To solve the distance task with a first-order computation, i.e., with linear receptive fields and without demodulation, would involve horrendous complications. The linear receptive fields needed for distance judgments are dumbbell-shaped receptive fields with one end of the dumbbell in each patch. Linear receptive fields are inherently phase sensitive with responses that vary from negative to zero to positive depending on just where in the receptive field the stimulus patches fall. Receptive fields would have to be duplicated for all orientations, distances, pairs of frequencies, and for at least four pairs of spatial phases (sin, cosine in each patch). Otherwise, for example, distance judgments would be impossible if the two bars being judged were of different spatial frequencies. In fact, the distance between two grating patches of different spatial frequencies is judged as accurately as the other distances (Burbeck, 1988). Demodulation resolves all these problems of first-order computations at once by transposing distance judgments to the lowest common domain.

For closely spaced bars, there is a significant difference in judging distance between same-contrast and opposite-contrast patterns. Again, there is the telltale *rectification-times-size interaction* that indicates two processing regimes: rectification dominating at large retinal sizes, and direct computation at small retinal sizes.

5.3.2. Bisection. Klein and Levi (1985) used a central line to bisect a spatial interval defined by two horizontal flanking lines. The flanking lines were either directly above and below the central line or displaced sideways. Observers judged which interval was smaller. With large retinal images, the sideways displacement was immaterial; with small retinal images, it was critical. This difference between results at close spacings and far spacings of lines in psychophysical judgments led Klein and Levi (1985) to postulate two regimes of detection mechanisms. They proposed a regime for small-size computations that relied on efficient linear filters (direct computation), but they did not propose a specific regime for large-size computations. However, the failure of the first-order small-size computation to account for the large-size results is consistent with the rectification proposal, although Klein and Levi's results do not specifically require rectification.

5.4. *Two processing regimes: Conclusions*

For bandpass filtered objects, different computations will be carried out depending on whether the object can be coded by neighboring sensors or whether it requires the coordination of information from distantly separated sensors. Nearest neighbor computations can use linear filters and can be highly efficient (first-order computations). Distant computations require demodulation (which is carried out by fullwave rectification) and information that is coordinated at higher levels of the computational hierarchy (second-order computations). The second-order computations, because they use rectification for demodulation, sacrifice statistical efficiency (impaired compared to ideal detectors) for computational simplicity (improved relative to attempting the computation at the same hierarchical level). Critical issues remain unresolved: How do first- and second-order computations combine to determine perception? To what extent are the noise sources associated with first- and second-order computations shared or independent?

It seems obvious that counting and labeling (rectification) operations will predominate over linear processes at higher perceptual and cognitive levels of processing. The surprise has been that simple rectification occurs so early in processing, being involved in retinal gain control and in the earliest stages of motion and pattern analysis. Presumably the appearance of rectification early in visual processing is determined by two factors: its economy of neural connectivity in a hierarchically organized nervous system and its ecological adequacy in our natural environment.

Acknowledgments

This work was supported by USAF, Life Sciences Directorate, Visual Information Processing Program, Grants 85-0364 and 88-0140.

REFERENCES

Adelson E H and Bergen, J. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am.* A 2, 284–299.

Anstis, S. M. (1970). Phi movement as a subtraction process. *Vision Res.* 10, 1411–1430.

Banks, M. S., Geisler, W. S. and Bennett, P. J. (1987). The physical limits of grating visibility. *Vision Res.* 27, 1915–1924.

Barlow, H. B. (1956). Retinal noise and absolute threshold. *J. Opt. Soc. Am.* 46, 634–639.

Barlow, H. B. (1957). Increment thresholds at low intensities considered as signal/noise discrimination. *J. Physiol.* 136, 469–488.

Berliner, J. E. and Durlach, N. I. (1973). Intensity perception. IV. Resolution in roving-level discrimination. *J. Acoust. Soc. Am.* 61, 1677–1685.

Braddick, O. (1974). A short-range process in apparent motion. *Vision Res.* 14, 519–527.

Burbeck, C. A. (1987). Position and spatial frequency in large-scale localization judgments. *Vision Res.* 27, 417–427.

Burbeck, C. A. (1988). Large-scale relative localization across spatial frequency channels. *Vision Res.* 28, 847–859.

Carlson, C. R. and Klopfenstein, R. W. (1985). Spatial-frequency model for hyperacuity. *J. Opt. Soc. Am.* A 2, 1747–1751.

Cavanagh, P. (1989). Motion: The long and the short of it. *Spatial Vision*, 4, 103–129.

Chubb, C. and Sperling, G. (1987). Drift-balanced random stimuli: A general basis for studying non-Fourier motion perception. *Invest. Ophthalmol. Visual Sci.* 28, 233.

Chubb, C. and Sperling, G. (1988a). Processing stages in Non-Fourier Motion Perception. *Invest. Ophthalmol. Visual Sci.* 29, 266.

Chubb, C. and Sperling, G. (1988b). Drift-balanced random stimuli: A general basis for studying non-Fourier motion perception. *J. Opt. Soc. Am. A* 5, 1986–2006.

Chubb, C. and Sperling, G. (1989a). Second-order motion perception: Space-time separable mechanisms. *Proceedings: 1989 IEEE Workshop on Motion*. IEEE Computer Society Press, Washington, DC, pp. 126–138.

Chubb, C. and Sperling, G. (1989b). Two motion perception mechanisms revealed by distance driven reversal of apparent motion. *Proc. Natl. Acad. Sci. USA* 86, 2985–2989.

Chubb, C. and Sperling, G. (1990). Texture quilts: Basic tools for studying motion from texture. *J. Math. Psychol.* 34, in press.

De Lange, H. Dzn (1954). Relationship between critical flicker-frequency and a set of low frequency characteristics of the eye. *J. Opt. Soc. Am.* 44, 380–389.

Dosher, B. A., Landy, M. S. and Sperling, G. (1989). Kinetic depth effect and optic flow: I. 3D shape from Fourier motion. *Vision Res.* 29, in press.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. Wiley, New York.

Gessler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychol. Rev.* 96, in press.

Grossberg, S. and Mingolla, E. (1985). Neural dynamics of perceptual grouping: Textures, boundaries, and emergent segmentations. *Percept. Psychophys.* 38, 141–171.

Ives, H. E. (1922). A theory of intermittent vision. *J. Opt. Soc. Am.* 6, 343–361.

Jamar, H. H. T., Campagne, J. C. and Koenderink, J. J. (1982). Detectability of amplitude- and frequency-modulation of suprathreshold sine-wave gratings. *Vision Res.* 22, 407–416.

Jamar, H. H. T. and Koenderink, J. J. (1985). Contrast detection and detection of contrast modulation for noise gratings. *Vision Res.* 25, 511–521.

Jesteadt, W., Wier, C. C. and Green, D. M. (1977). Intensity discrimination as a function of frequency and sensation level. *J. Acoust. Soc. Am.* 61, 169–177.

Klein, S. and Levi, D. M. (1985). Hyperacuity thresholds of 1 sec: Theoretical predictions and empirical validation. *J. Opt. Soc. Am. A* 2, 1170–1190.

Knutsson, H. and Granlund, G. H. (1983). Texture analysis using two-dimensional quadrature filters. *IEEE Computer Society Workshop on Computer Architecture for Pattern Analysis and Image Database Management (CAPAIDM)*. IEEE Computer Society, Silver Spring, MD, pp. 206–213.

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *J. Neurophysiol.* 16, 37–68.

Landy, M. S., Sperling, G., Dosher, B. A. and Perkins, M. (1987). From what kind of motions can structure be inferred? *Invest. Ophthalmol. Visual Sci. (Suppl.)* 28, 233.

Legge, G. E. and Foley, J. M. (1980). Contrast masking in human vision. *J. Opt. Soc. Am.* 70, 1458–1471.

Legge, G. E., Kersten, D. and Burgess, A. E. (1987). Contrast discrimination in noise. *J. Opt. Soc. Am.* A 4, 391–404.

Lelkens, A. M. M. and Koenderink, J. J. (1984). Illusory motion in visual displays. *Vision Res.* 24, 1083–1090.

MacMillan, N. A. (1987). Beyond the categorical/continuous distinction: A psychophysical approach to processing modes. In: *Categorical Perception*, S. Harnad (Ed.). Cambridge University Press, New York, pp. 53–85.

Marr, D. (1982). *Vision, A Computational Investigation into Human Representation and Processing of Visual Information*. W. H. Freeman and Company, San Francisco.

Nakayama, K. and Silverman, G. (1985). Detection and discrimination of sinusoidal grating displacements. *J. Opt. Soc. Am. A* 2, 267–274.

Pantle, A. and Picciano, L. (1976). A multistable movement display: Evidence for two separate motion systems in human vision. *Science* 193, 500–502.

Pantle, A. and Turano, K. (1986). Direct comparisons of apparent motions produced with luminance, contrast-modulated (CM), and texture gratings. *Invest. Ophthalmol. Visual Sci.* 27, 141.

Parish, D. H. and Sperling, G. (1987a). Object spatial frequencies, retinal spatial frequencies, and the efficiency of letter discrimination. *Mathematical Studies in Perception and Cognition, 87-8*, Department of Psychology, New York University, Pp. 30.

Parish, D. H. and Sperling, G. (1987b). Object spatial frequency, not retinal spatial frequency, determines identification efficiency. *Invest. Ophthalmol. Visual Sci. (Suppl.)* 28, 339.

Pavel, M., Sperling, G., Riedl, T. and Vanderbeek, A. (1987). The limits of visual communication: The effect of signal-to-noise ratio on the intelligibility of American Sign Language. *J. Opt. Soc. Am. A* 4, 2355–2365.

Pelli, D. G. (1981). *Effects of Visual Noise*. Ph.D. dissertation, University of Cambridge, Cambridge, England.

Pearday, K. (1987). Three-dimensional structure from long-range apparent motion. *Perception*, 15, 619–625.

Ramachandran, V. S., Madhusudhan Rao and Vidyasagar, T. R. (1973). Apparent movement with subjective contours. *Vision Res.* 13, 1399–1401.

Reichardt, W. (1957). Autokorrelationsauswertung als Funktionsprinzip des Zentralnervensystems. *Z. Naturforsch.* 12b, 447–457.

Ricco, R. R. (1928). Differential intensity sensitivity of the ear. *Phys. Rev.* 31, 867–875.

Robinson, D. W. and Dadson, R. S. (1956). A redetermination of the equal-loudness relations for pure tones. *Br. J. Appl. Phys.* 7, 166–181.

Schade, O. H. (1952). Optical and photoelectric analog of the eye. *J. Opt. Soc. Am.* 46, 721–739.

Scharf, B. and Buus, S. (1986). Audition I. Stimulus, physiology, thresholds. In: *Handbook of Perception and Performance, Vol. I*. K. Boff, L. Kaufman, and J. Thomas (Eds.) Wiley, New York, pp. 14-1 to 14-71.

Shapley, R. and Enroth-Cugell, C. (1984). Visual adaptation and retinal gain control. *Prog. Retinal Res.* 3, 263–266.

Sperling, G. (1976). Movement perception in computer-driven visual displays. *Behavior, Res. Methods Instrum.* 8, 144–151.

Sperling, G. and Chubb, C. C. (1987). Non-Fourier motion and texture perception. Paper presented at *AFOSR Program in Visual Information Processing*, Annapolis, MD, December 16, 1987.

Sperling, G. and Parish, D. H. (1985). Forest-in-the-trees illusions. *Invest. Ophthalmol. Visual Sci. (Suppl.)* 26, 285.

Sperling, G. and Sondhi, M. M. (1968). Model for visual luminance discrimination and flicker detection. *J. Opt. Soc. Am.* 58, 1133–1145.

van Nes, F. L. and Bouman, M. A. (1967). Spatial modulation transfer in the human eye. *J. Opt. Soc. Am.* 57, 401–406.

van Santen, J. P. H. and Sperling, G. (1984). Temporal covariance model of human motion perception. *J. Opt. Soc. Am. A* 1, 451–473.

van Santen, J. P. H. and Sperling, G. (1985). Elaborated Reichardt detectors. *J. Opt. Soc. Am. A* 2, 300–321.

Victor, J. D. and Conte, M. M. (1989a). Cortical interactions in texture processing: scale and dynamics. *Visual Neurosci.* in press.

Victor, J. D. and Conte, M. M. (1989b). Motion mechanisms have only limited access to form information. Manuscript.

Wallach, H. and O'Connell, D. N. (1953). The kinetic depth effect. *J. Exp. Psychol.* 45, 205–217.

Watson, A. B. and Ahumada, A. J. Jr. (1983). A linear motion sensor. *Perception*, 12, A17.

Watt, R. J. and Morgan, M. J. (1985) A theory of the primitive spatial code in human vision. *Vision Res.* 25, 1661–1674.

Westheimer, G. and McKee, S. P. (1977) Perception of temporal order in adjacent visual stimuli. *Vision Res.* 17, 887–892.

NOTES

1. Informal observations. Variations in contrast and luminance were not reported in Parish and Sperling (1985a).
2. The nomenclature was suggested by P. Cavanagh (1990).
3. Although it is embedded in a much more complex framework, Grossberg and Mingolla (1985) incorporate a full-wave rectifying stage in their general model of texture and boundary perception that appears to deal with second-order stimuli. However rectification and similar nonlinear operations such as squaring do not, in and of themselves, imply second-order processing. For example, Adelson and Bergen's (1985) detector of directional motion energy is equivalent to the Reichardt motion model (van Santen and Sperling, 1984, 1985) and to Knutsson and Granlund's (1983) texture-orientation model. All of these models embody a nonlinear squaring stage (or the equivalent) and they merely perform first-order computations; none can detect second-order motion or orientation.
4. Ives (1922) anticipated subsequent linear theories of visual threshold phenomena but he was ignored by the psychophysicists of his time because they did not understand linear systems theory.

Vision Res. Vol. 29, No. 12, pp. 1789–1813, 1989
Printed in Great Britain

0042-6989/89 $3.00 + 0.00

KINETIC DEPTH EFFECT AND OPTIC FLOW—I. 3D SHAPE FROM FOURIER MOTION

BARBARA A. DOSHER,[1] MICHAEL S. LANDY[2] and GEORGE SPERLING[2]

[1]Psychology Department, Box 28, Schermerhorn Hall, Columbia University, New York, NY 10027 and [2]Psychology Department, New York University, Washington Square, New York, NY 10012, U.S.A.

(*Received* 17 *August* 1988; *in revised form* 10 *February* 1989)

Abstract—Fifty-three different 3D shapes were defined by sequences of 2D views (frames) of dots on a rotating 3D surface. (1) Subjects' accuracy of shape identifications dropped from over 90% to less than 10% when either the polarity of the stimulus dots was alternated from light-on-gray to dark-on-gray on successive frames or when neutral gray interframe intervals were interposed. Both manipulations interfere with motion extraction by spatio-temporal (Fourier) and gradient *first-order* detectors. *Second-order* (non-Fourier) detectors that use full-wave rectification are unaffected by alternating-polarity but disrupted by interposed gray frames. (2) To equate the accuracy of two-alternative forced-choice (2AFC) planar direction-of-motion discrimination in standard and polarity-alternated stimuli, standard contrast was reduced. 3D shape discrimination survived contrast reduction in standard stimuli whereas it failed completely with polarity-alternation even at full contrast. (3) When individual dots were permitted to remain in the image sequence for only two frames, performance showed little loss compared to standard displays where individual dots had an expected lifetime of 20 frames, showing that 3D shape identification does not require continuity of stimulus tokens. (4) Performance in all discrimination tasks is predicted (up to a monotone transformation) by considering the quality of first-order information (as given by a simple computation on Fourier power) and the number of locations at which motion information is required. Perceptual first-order analysis of optic flow is the primary substrate for structure-from-motion computations in random dot displays because only it offers sufficient quality of perceptual motion at a sufficient number of locations.

Kinetic depth effect Structure from motion Shape identification Fourier motion

INTRODUCTION

A sequence of 2D projected images (frames) of a moving 3D object is sometimes perceived as a moving 3D shape. When each isolated 2D frame is uninformative about 3D shape, but the sequence causes a 3D shape to be perceived, this is called the *kinetic depth effect*, after Wallach and O'Connell (1953). When a computer algorithm recovers 3D shape from a 2D frame sequence, it is called *structure from motion* (Ullman, 1979).

There are two classes of proposed models for deriving 3D shape from 2D frame sequences; we designate them as *feature-correspondence models* and *flow-field models*.

Feature-correspondence models

Feature-correspondence models use geometric constraints, usually coupled with assumptions of rigidity, to derive shape. Examples of algorithms that derive a 3D configuration from a set of n points (or similar features) displayed in each of m frames are Hoffman and Bennett (1985) and Ullman (1979, 1985), or see Braunstein, Hoffman, Shapiro, Andersen and Bennett (1987) for a more empirical treatment. A list of visual features is identified and located in 2D space on each frame. In this class of model, the correspondence of point n in frame m with equivalent point n in frame $m+1$ is assumed to be known. Using Euclidean geometry and the assumption of object rigidity, a 3D location for each feature on each frame is derived. The set of 3D locations determines object shape

Flow-field models

Flow-field models derive object shape from local velocity information described by optic flow fields. An object is described by many points or other features densely scattered on its surface and possibly throughout its volume. The flow-field is computed from the velocities of groups of points over a sequence of frames Flow-field velocities determine relative depths and orientations and thereby object shape (e.g

Atch- 16

Clocksin, 1980; Hoffman, 1982; Koenderink & van Doorn, 1986). Flow-field models suggest that a sequence of frames might be considered not as an abstract list of features with associated location information, but as a motion stimulus to one or more motion-detection mechanisms. In this article, we are primarily concerned with determining the nature of this motion stimulus.

FIRST-ORDER AND SECOND-ORDER MOTION SYSTEMS

We consider here three kinds of motion-detectors: two first-order detectors, which we designate as (1) spatio-temporal motion energy detectors and (2) gradient detectors, and (3) second-order detectors. A first-order detector detects motion in stimuli that would yield motion to a local spatio-temporal Fourier analysis; a second-order detector may detect such motion but also detects motion in a wide class of stimuli that do not yield directional motion under any kind of Fourier analysis. We examine these kinds of detectors in more detail below.

Fourier motion-energy detectors: the elaborated Reichardt detector (ERD)

Low-level motion mechanisms are now thought to be based on systems that approximate a local spatio-temporal Fourier analysis of frame sequences (Adelson & Bergen, 1985; van Santen & Sperling, 1985; Watson & Ahumada, 1983; Watson, Ahumada & Farrell, 1986). Indeed, whenever the spatio-temporal frequency components of a stimulus differ in temporal frequency, the output of these mechanisms is simply the sum of their responses to the individual spatio-temporal Fourier components of the stimulus (derived from their equivalence to Reichardt detectors—van Santen & Sperling, 1984a, b). The Reichardt detector (Reichardt, 1957) was the first computational motion detector. The elaborated Reichardt detector (van Santen & Sperling, 1984a, b, 1985) successfully extended the basic scheme to the prediction of human psychophysical data, although there were earlier attempts (e.g. Foster, 1969, 1971). The motion models of Watson and Ahumada (1983) (when elaborated) and of Adelson and Bergen (1985) have motion-detection mechanisms that are defined differently but have been shown to be equivalent to Reichardt detectors at their final outputs (van Santen & Sperling, 1985), although the order of intermediate operations is different.

Motion discrimination (e.g. the discrimination of leftward from rightward motion) now appears to be a different process than velocity discrimination. The elaborations of the basic motion-detection mechanism to account for velocity discrimination are quite complex (e.g. Watson & Ahumada, 1985; Heeger, 1987) and involve the interplay of many elementary motion detectors. Since all these models ultimately depend on a basic mechanism that is equivalent to an elaborated Reichardt detector (ERD), we shall describe the ERD in more detail.

A Reichardt motion detector consists of two component half-detectors. One half-detector compares the intensity at point A, time t with the intensity at point B, time $t + \Delta t$ (see Fig. 1). The other half-detector looks at (B,t) and

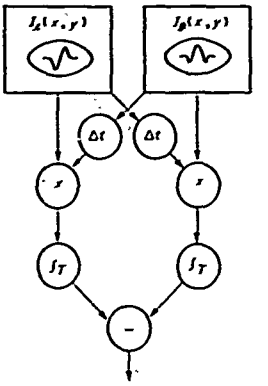


Fig. 1. A schematic illustration of an elaborated Reichardt detector (van Santen & Sperling, 1985), one implementation of a spatio-temporal motion analyzer. Image intensity at location A at time t is correlated (multiplied) by image intensity at location B at time $t + \Delta t$ (left half-detector). Similarly, image intensity at location B at time t is correlated (multiplied) by image intensity at location A at time $t + \Delta t$ (right half-detector). These correlation values are temporally integrated over some time domain T, and compared (subtracted) to yield a direction-of-motion signal for that detector. Orientation and velocity tuning are determined by the selection of receptive fields I_A and I_B and Δt. Spatial scale is determined by the spatial function which senses image intensity. Outputs of populations of such detectors of various scales, locations, and velocity tuning must be integrated with subsequent decision rules. Further elaborations are required to construct velocity sensors.

$(A,t+\Delta t)$. While each half-detector can detect motion by itself, the two together have some important advantages. They signal motion in opposite directions by outputs of opposite sign, and by canceling evidence for movement in opposite directions, they help to disambiguate flicker and other nonmotion stimuli from true motion.

To account for psychophysical data, the spatial points A and B are replaced with spatio-temporal receptive fields, I_A and I_B, and the pure delay Δt is replaced with a linear filter. The receptive fields I_A and I_B determine the spatial orientation-tuning of the detector, and I_A and I_B taken with the time delay Δt jointly determine the velocity tuning. Theories of human motion perception which we have discussed assume that populations of such detectors exist in different sizes (scales) and at each scale they are tuned to different orientations and velocities. The aggregated outputs of all these detectors are combined by a *voting* (decision) rule to predict the direction of perceived motion at each spatial location and time.

ERDs (and hence the various equivalent spatio-temporal motion-energy models) account for a wide variety of critical data on direction of motion discrimination (van Santen & Sperling, 1984a, 1985). To provide velocity sensing, outputs of arrays of basic spatio-temporal motion detectors must be combined (Watson & Ahumada, 1985; Heeger, 1987), because an isolated ERD will not function adequately as a velocity detector. Stimulus contrast and many factors relating to velocity tuning are confounded in the response of any one motion detector. Watson and Ahumada (1985) propose direct coding of the temporal frequency of sets of motion detectors. Heeger (1987) compares the overall pattern of responses of a set of motion detectors to an unknown stimulus to the patterns produced by known training stimuli.

Gradient detectors

A second class of first-order motion detection mechanisms uses gradients in the computation. Examples are Limb and Murphy (1978), Fennema and Thompson (1979), Horn and Schunk (1981), Marr and Ullman (1981), and Harris (1986). Basically, these models find local areas where luminance $I(x,y,t)$ varies as a function of (x,y), i.e. has a nonzero spatial gradient $\nabla I(x,y,t) \neq 0$. The velocity v is determined by the ratio of the change in $I(x,y,t)$ as a function of time to the change in $I(x,y,t)$ as a function of space. Gradient models do a single local computation that embraces both the Reichardt motion detection mechanism and the subsequent velocity stage of the flow-field models.

Whenever the spatial luminance gradient is small, velocity estimates are extremely unstable. Therefore, Adelson and Bergen (1986) proposed weighting the local velocity estimates by a "confidence" value. Choosing the "confidence" level as the local value of the squared gradient converts the gradient computation into a least-squares estimate of velocity (Lucas & Kanade, 1981), a computation that can be carried out by the first-order motion-energy/elaborated-Reichardt systems that we outlined above. Thus, while at first glance gradient computations seem quite different from Fourier first-order computations, the difference vanishes when a realistic gradient computation is made (Adelson & Bergen, 1986).

Second-order motion detection

Stable perception of direction of movement and of velocity can arise from complex stimuli which are essentially invisible to first-order motion detectors—they fail to report any consistent direction (Chubb & Sperling, 1988a, b). Motion detectors to perceive Chubb and Sperling's motion stimuli require two stages of linear filtering separated by a full-wave rectification stage that computes the absolute value of contrast. For the present stimuli, however, the linear filtering stages are unnecessary and will be omitted. Because of the necessity of a two-stage analysis (first rectification with or without filtering, then Reichardt-or-equivalent motion detection), motion detectors that can detect such stimuli are called *second-order*. Early evidence (Chubb & Sperling, 1987) suggests that second-order systems may operate primarily foveally and with lower spatial resolution than first-order detectors. Since they depend on rectification, with inevitable loss of information, second-order systems have higher contrast thresholds than first-order systems (Chubb & Sperling, 1989a, b).

First-order and second-order systems and KDE

This paper asks whether the *ability* of humans to perceive 3D shape from a 2D frame sequence depends on the strength of evidence supplied to first-order motion mechanisms. This question stands in sharp contrast to much of the historic work on kinetic depth effect, which emphasized cues such as perspective (e.g. Braunstein, 1962),

numerosity (Green, 1961), or occlusion (Andersen & Braunstein, 1983) and their effect on the *nature* of a shape percept. We ask whether strong input to a first-order motion system is necessary to support shape perception. Our strategy is to introduce factors such as flicker or contrast (polarity) reversal that weaken or disrupt a first-order motion mechanism. We can then ask whether the ability to perceive 3D shape is especially degraded. Symmetrically we ask, do second-order systems support 3D shape perception?

In the experiments of this paper, kinetic depth displays are rendered as dots scattered randomly on a 3D surface. These are projected as a 2D image of bright dots on a neutral gray background. Figure 2a schematically illustrates spatio-temporal analysis of a moving intensified (brighter) dot on a gray background. A frame sequence defines the stimulus as a function in (x,y,t), where x and y represent locations in the picture plane, and t represents frames (time). Figure 2 simplifies the analysis by showing only the (x,t) plane. A line in the (x,t) plane represents the x-component of velocity. A spatio-temporal receptive field here tuned to precisely the velocity of the illustrated points is a core component of one representational form of the Fourier energy motion detectors (Adelson & Bergen, 1985; Watson & Ahumada, 1984; and by equivalence, the ERD, van Santen & Sperling, 1984a, b).

Figure 2b illustrates a manipulation which interspersses gray frames between motion samples, but maintains the same velocity. This reduces the amplitude of the fundamental motion component by half and introduces many low-amplitude motion components opposite in direction to the fundamental. One such opposite direction detector is illustrated in Fig. 2b. An alternating gray frame display is equivalent to a half-wave rectification of a polarity alternation stimulus (see below). For our gray-frame stimuli, the total Fourier energy in each direction is approximately equal. If the sensitivities to the various spatio-temporal motion components were equal, the energy in each direction would balance and neutralize the Fourier system. Empirically, at constant velocity, reducing the number of samples (as in a gray frame versus a standard motion stimulus) always impairs the perceived quality of stroboscopic motion (Sperling, 1976). Reducing blank (background level) interstimulus intervals to about 20 msec (and hence varying velocity) improves planar apparent motion between two alternating frames of random dots (Braddick, 1973, 1974) or multi-frame sequences (Burt & Sperling, 1981).

Figure 2c illustrates a motion stimulus which alternates polarity of the motion token between intensities higher and lower than the neutral (mean) gray level. Polarity alternation provides cancelling inputs to local spatio-temporal filters

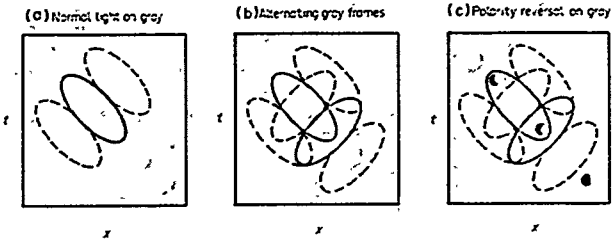


Fig 2. (a) Schematic illustration of a simple spatio-temporal sensor operating on a moving white dot on a gray background. One dimension of space x, and time t are represented. The center (solid ellipse) has a weight of $+1$; each of the flanks (dotted ellipse) has a weight of $-\frac{1}{2}$. The geometry and orientation of the hypothetical receptive field represent the preference for a particular spatial scale, direction, and velocity. (b) Same sensor as (a) operating on a stimulus with interleaved gray frames, and a second sensor sensitive to the opposite velocity. The magnitude of the stimulation of the center of sensor 1 equals the combined magnitude of the stimulation of the two flanks of sensor 2. At this scale, there is equal evidence for both orientations, i.e., both velocities. (c) Same sensor as (a) operating on a stimulus with tokens alternating polarity above and below the gray background level. Sensor 1 receives oppositely signed inputs in its center and has a weak output. Sensor 2 receives inputs in its surround opposite in sign from those in its center and therefore has a large output. Alternating polarity yields strong evidence for orientation from upper right to lower left, i.e. for motion opposite to the direction in (a)

tuned to the "veridical" motion direction; alternation, as illustrated, stimulates large-scale detectors tuned to the opposite direction (Anstis, 1970; Anstis & Rogers, 1975; Chubb & Sperling, 1988b; Rogers & Anstis, 1975). Like the spatiotemporal energy models, the gradient methods, which examine changes in luminance patterns over time, are also disrupted by polarity reversal.

We investigate interspersed gray frames and polarity reversal (and other manipulations; see Landy, Dosher, Sperling & Perkins, 1988) that may disrupt first-order processes. We determine whether 3D shape extraction is disrupted. It is also important to determine whether any such disruption is special to 3D shape extraction processes, or whether it can be accounted for exactly by decrements in simpler 2D visibility and motion tasks.

The objective measure of 3D shape recovery

The essence of kinetic depth perception is the addition of depth information to a 2D image to create a perception of a 3D object shape. We ask whether kinetic depth percepts depend on first-order motion analysis. In order to have more than a qualitative answer to this question, it was first necessary to develop an objective index of 3D shape perception. To this end, we (Sperling, Landy, Dosher & Perkins, 1989) developed a shape identification task with a very low guessing baserate (near 2%) and a large performance range (up to 95+%). This task requires subjects to identify a display as depicting one of a large lexicon (53) of three-dimensional (3D) surface shapes. In this paper, we also use comparison tasks such as detection, direction discrimination and motion segmentation in several control studies.*

GENERAL METHODS

Apparatus

Stimuli were pre-generated and stored on a Vax 11/750 computer that shipped images to an Adage RDS-3000 image display system. A Conrac 7211C19 RGB color monitor was used for display, operating at a refresh rate of 60 Hz, noninterlaced. Only the green beam of the monitor was used.

Procedure

Displays were seen through a viewing tunnel and circular aperture, which provided monocular viewing at a viewing distance of 1.6 m. The circular aperture was slightly larger than the displays. The size, intensity, timing and content of the displayed frame sequences are listed below for each experiment separately. Following each display sequence, the subject pressed keys or typed the required judgement. The primary task was shape identification. Control tasks included standard two-interval detection, direction-of-motion discrimination, and motion segmentation. Displays were viewed in mixed lists within experiments.

The methods sections for Expts 1–6 are presented together below, in the order in which the results will be discussed. This allows an uninterrupted presentation of the arguments in the Results section, where motivation for the particular conditions and experiments can be found. The experiments were actually run in the following order: 1, 3, 5, 2, 6 then 4.

The displays, or conditions, for Expts 1–3—the 3D shape identification experiments—are summarized in Table 1. The displays, or conditions, for Expts 4–6—planar motion experiments—are summarized in Table 2. Distinct display types are numbered continuously in the two tables.

METHOD: EXPERIMENT 1 (MAIN)

Identification stimuli

The main experiment compared objective performance levels on standard kinetic depth displays with performance on comparable displays that disturb or weaken first-order motion cues. The objective measure was percent correct identification. The shape lexicon was based on peaks, valleys, and flat regions located in one of two triangular layouts. Figure 3a shows the two triangular layouts on a square ground, and Fig. 3b shows some examples of shapes. Fig 3c illustrates a shape movement, and Fig. 3d indicates the size of a single display frame. Stimulus identification consisted of reporting the layout (Up vs Down), the sign of the bump (+ = peak, 0 = flat, − = valley) in each of locations 1, 2, and 3, and the direction of rotation. (See Sperling et al., 1989, for details.)

For the 3D shape identification task, feedback consisted of a list of the correct responses.

*Preliminary reports of these experiments are contained in Landy, Sperling, Dosher and Perkins (1987), Landy, Sperling, Perkins and Dosher (1987) and Dosher, Landy and Sperling (1988).

Table 1. Display types for Expts 1–3

Display	Motion cue[a]	Density cue[b]	Rotation speed[c]	Intensity ± increments[d]	Dot lifetime[e]
	Task: large lexicon shape identification				
Experiment 1 (Main)					
1. With density	3D	Y	Standard	1:1	≤30
2. Standard	3D	N	Standard	1:1	≤30
3. With density	3D	Y	Half	1:1	30
4. Standard	3D	N	Half	1:1	≤30
5. Alternating polarity	3D	N	Standard	1:−1	≤30
6. Alternating polarity	3D	N	Standard	0.5, −0.5	≤30
7. Alternating gray	3D	N	Half	1:0	≤30
8. Alternating gray	3D	N	Standard	1:0	≤30
9. Alternating contrast	3D	N	Standard	2:1	≤30
10. Alternating contrast	3D	N	Standard	1.5:0.5	≤30
11. Density only	Random	Y	Standard	1:1	1
Experiment 2 (Equated contrast)					
12. Standard	3D	N	Standard	V:V	≤30
Experiment 3 (Lifetimes)					
2. Standard	3D	N	Standard	1:1	≤30
13. 3-Frame	3D	N	Standard	1:1	3
14. 2-Frame	3D	N	Standard	1:1	2

[a]3D motion cues refers to 2D projections of 3D moving stimuli. Random refers to random motion correspondences arising from uncorrelated new dot samples on each frame.

[b]Dot-density cues removed by minimal (<5%) dot scintillation.

[c]Standard rotation speed: was ±25 deg sinusoidal rotation per 30 new frames, 15 new frames per sec with 4 sync cycles per new frame. Half rotation speed: ±25 deg sinusoidal rotation per 30 new frames; 7.5 new frames per sec with 8 sync cycles per new frame (conditions 3, 4) or 15 new frames per sec with 4 sync cycles per new frame (condition 7) (see text).

[d]The numbers code the increments or decrements in intensification of dots on a neutral gray background. 1 refers to 1 × the standard increment level, and −1 refers to 1 × the standard decrement level. The value to the left of the colon refers to dot intensification on odd frames; the value to the right to even frames. For example, 1:1 means dots received the same standard increments on all frames; 1:0 means dots received standard intensification on odd frames, and no intensification on even frames; etc. Gray background was between 31 and 38 cd m^2. Standard increments (and decrements) were between 13 and 21 extra (or fewer) μcd per dot. See the text for exact values for each subject. The value V refers to fraction (<1) of standard increment intensity which equates non-alternating stimuli to alternating polarity stimuli for percent correct planar motion direction judgements (see Expt 5). Intensities for V were between approximately 0.5–0.6, or between 8 and 10 μcd per dot.

[e]Lifetime refers to the number of new frames that the same dots on the 3D surface appear in during the stimulus sequence. Since the display sequences were 30 new frames long, a lifetime of 30 frames is maximal. The value ≤30 refers to nominal lifetime of 30 frames, subject to scintillation for density control. Conditions (13) and (14) resample one third and one half of the dots in the stimulus per frame, respectively, yielding scintillation values of 33% and 50%.

For any stimulus, there were two correct responses, which are depth-reversals of one another; the depth reversals are coupled with opposite perceived directions of rotation. Subjects were initially shown perspective drawings of shapes and instructed in naming performance. Subjects were trained in practice sessions until they achieved approximately 85% correct on the easiest stimuli.

The standard kinetic depth display consisted of white dots on a mid-intensity (gray) background. The displays were 300 dot random subsamples of the picture plane, displayed with an x,y resolution of 182 × 182 pixels.* Projections were parallel. Peaks or valleys had simulated height equal to half the side of the square ground. The smooth surface was constructed by smoothing of a spline interpolation over the stimulus peaks and the ground. The surface was initially parallel to the projection plane, and rotated first right (or left) 25 deg, back through to left (or right) by 25 deg, and then back full-forward (25 deg amplitude sinusoidal rotation) over a period of 30 new image frames. Stimulus edges never appeared in the display window. The displays assumed no occlusion of dots by the 3D surface (transparency). The standard display rate was 15 new frames (with changed frame contents) per second. Each new

*The number of dots actually varied slightly from 300 due to sampling of dots at or near the windowed edges

Table 2. Display types for Expts 4–6

Display	Motion cue[a]	Number of patches[b]	Motion direction	Intensity± increments[c]	Task
Planar motion experiments					
Experiment 4 (Visibility)					
15–19. Standard	2D	1	L or R	5 levels	Detection (2IFC)
20–24. Alternating polarity	2D	1	L or R	±5 levels	Detection (2IFC)
Experiment 5 (Motion direction)					
25–29. Standard	2D	1	L or R	5 levels	Direction (2AFC)
30–34. Alternating polarity	2D	1	L or R	±5 levels	Direction (2AFC)
Experiment 6 (Motion segmentation)					
35. Standard	2D	9	8L/1R or 1L/8R	1:1	Odd motion (9AFC)
36. Alternating polarity	2D	9	8L/1R or 1L/8R	1:−1	Odd motion (9AFC)

[a]2D motion cue refers to uniform field motion of a random dot field in a larger background of neutral gray or of dynamic random dot noise. Planar motion was 1 pixel per new frame, 15 new frames per sec, or 4 sync cycles per new frame. See text for details.
[b]Patches were 48 × 48 pixels. Single patches were embedded in a larger background. The 9-patch displays were arranged in a 3 × 3 square grid.
[c]Dots were displayed as increments or decrements on a gray background. The intensities were varied as percentages of the standard increments and decrements, which are labeled as in Table 1. Variable intensity increments differed across subjects (see text).

frame was shown for 4 sync cycles, at a monitor sync rate of 60 Hz. Half speed displays either showed new frames every 8 sync cycles, or at 4 sync cycles with interleaved gray frames. In the data of Sperling et al. (1989), a similar white-on-black display condition yielded identification performance in the 95% range. Other conditions modified this standard display.

Display geometry and timing

The 3D shape display was confined to the central 182 × 182 pixels of a 512 × 512 raster (60 Hz, no interlace). Background luminance was uniform over the entire 512 × 512 area. The 182 × 182 display area subtended 3.7 by 4.2 deg at a viewing distance of 1.6 m that was controlled by viewing tube and aperture. On each trial, a fixation spot appeared for 1 sec, followed by 1 sec of blank (gray) screen, then the rotating stimulus for 2 sec (4 sec for half-speed displays). The screen was blank until the next trial was initiated. Responses were typed into a separate keyboard, and feedback (correct stimulus identification) appeared on a separate CRT.

Calibrated intensities

The display monitor was calibrated to equate the light and dark dots on the gray background, i.e. the luminance energy gain of increments and the luminance energy loss of decrements.* Three subjects participated. For subject MSL, the standard intensity condition consisted of background luminance of 31.8 cd/m² (average of 11.6 μcd/pixel) with 13.2 μcd additional (or lowered) intensification for each stimulus *dot* (at viewing distnace of 1.6 m). For subject CFS, the background was 31.0 cd/m² (average of 11.3 μcd/pixel) with increments or decrements of 13.2 μcd/dot. For subject JBL, the background was 38.8 cd/m² (average of 14.2 μcd/pixel) with increments or decrements of

*The linearization of the monitor depended on the average intensification level. To equate light and dark dots required calibration on the same gray level, and with display conditions as closely related to the actual displays as possible. A regular grid of one in nine pixels was nominally assigned the dark intensity and the remaining pixels assigned the gray background level. The decrement (in cd m²) relative to a uniform field of background intensity was equated to the increment when one in nine pixels were assigned the light intensity on a gray background level. One in nine pixels is an approximation to the sparse displays of the actual stimuli, while still providing stable measurements with an UDT-161CRT photometer. The increment in intensification due to each stimulus dot (in μcd/dot) was computed from the field increment. Although a stimulus dot is nominally one pixel, our calibrations show that intensification affects neighboring pixels via the point spread function of the monitor and phosphor nonlinearities.

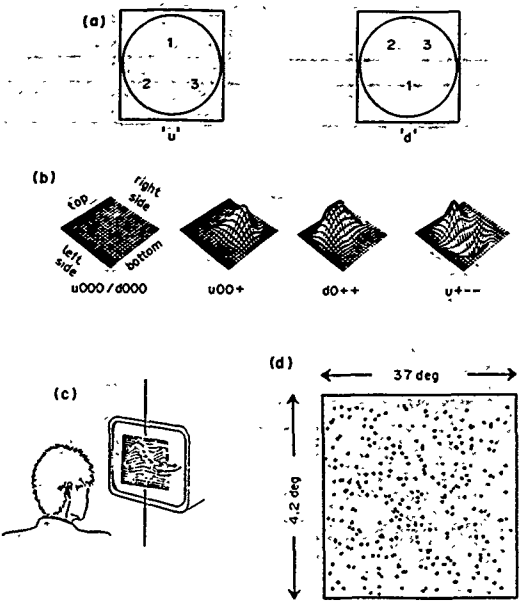


Fig. 3. (a) Illustration of the upward and downward pointing triangular layout of peak and valley locations in the shape lexicon. Members of the lexicon may have either the upward or downward layout, and either a peak, valley, or ground value at each of the three locations. (b) Examples of a number of shapes in the shape lexicon as defined by a rectangular grid spline over peaks and valleys. Actual stimuli consisted of parallel projections of dots sprinkled over these shapes undergoing sinusoidal rotation. Subjects were required to identify the shape and the direction of rotation. (c) Schematic illustration of the shape identification displays with rotation. (d) A single frame of a 2D image sequence for the shape identification task.

20.9 μcd/dot. Note: subject CFS could not be refracted completely to normal vision; his corrected Snellen acuity was approximately 20/40. All other subjects had normal or corrected-to-normal vision.

Conditions

The main experiment included 11 display conditions. Each of the 54 possible shape stimuli appeared once in each of the 11 conditions, for 594 identification trials per subject. All of these stimuli were shown in one large mixed list, divided over 4 sessions.

The relevant characteristics of the 11 display conditions are listed in Table 1. All displays in this experiment, except condition (11), depict the motion of 3D shapes in 2D projection. An unconstrained subsampling of points on the 3D shapes, includes density cues that result when peaks and valleys cause dots to bunch together in the projection of the 3D surface onto the 2D image plane. Except in displays in conditions (1), (3), and (13), subsampling of dots was constrained such that local density was constant across the display. Density cues were eliminated from the image sequences by adding or subtracting a small number of points on each frame so as to equate dot density within local regions comprising approximately 1/10 × 1/10th of the stimulus area. Constant-density subsampling introduced minor levels of apparent scintillation. The amount of scintillation can be expressed as the average percentage of dots *not* maintained from frame m to frame $m + 1$, or equivalently, in the expected lifetime of dots. Over all the density-controlled (no density cue) displays in the experiment, the average scintillation was 5%, yielding an expected dot lifetime

of 20 frames for the dots of frame 1. (These displays are indicated as ≤30 in Table 1.) Condition (11) extracts the local density cues in (1), but eliminates systematic motion information. Time- and position-dependent density is generated by random sampling from the rotating 3D shape with a new random sample for each frame (dot lifetime of 1 frame). This destroys systematic motion cues, but maintains local variations in dot density under rotation.

Most displays depict a standard rotation speed as described above. In conditions 3 and 4, half-speed rotation is produced by displaying each new frame for 8 (sync) repetitions (instead of 4 in the standard condition). The half-speed gray frame conditon (7) is accomplished by interleaving 4 repetitions of each new frame with 4 repetitions of gray frame. Full-speed gray frame condition (8) is accomplished by interleaving 4 repetitions of *every other* new frame of the standard stimulus with 4 repetitions of gray frame.

Standard displays depict the 3D shapes by displaying bright dots, of a selected standard intensity of increment on a neutral (gray) background. Intensity listings in Table 1 refer to a multiple of the standard dot intensification, positive for increments and negative for decrements. In alternating polarity displays, the dots are bright in odd frames, and dark on even frames (labelled 1: −1). In alternating gray displays, gray background is displayed on all even frames (labelled 1:0). Other non-standard increments serve as controls.

METHOD: EXPERIMENT 2 (EQUATED CONTRAST IDENTIFICATION)

Conditions

The task in this experiment was 3D shape identification; it was conducted with displays that had been equated for discrimination of motion direction by reducing dot intensity by an amount determined from Expt 5. Subjects viewed *standard* 3D shape identification displays—Table 1, condition (2)—in which the dot increments had been reduced (condition 12). The data for the standard (non-alternating condition) in Expt 5, by interpolation, allowed the selection of an increment intensity which would approximately equate the percent correct motion direction judgement of the standard condition with polarity alternation stimuli at full intensity increments and decrements. This equal-direction-discrimination value was determined separately for each of the two subjects. Each of the 54 identification stimuli was presented in random order.

Display geometry and calibrated intensities

Viewing conditions were the same as those described in Method Experiment 1. Calibrated intensities were: for MSL, the background intensity was 31.0 cd/m² with increment/decrement intensity of 8.8 μcd/dot. For JBL, the background intensity was 38.0 cd/m² and increment intensity was 9.6 μcd/dot.

METHOD: EXPERIMENT 3 (LIFETIMES)

Conditions

This experiment compared three conditions in which the lifetimes of the dots were 2 frames, 3 frames and ≤30 frames (continuous) (conditions 14, 13 and 2, respectively, under Expt 3 in Table 1). See Fig. 6a for an illustration. New dots were subsampled randomly, with additional subsampling to eliminate density cues for all conditions of this experiment. The task was 3D shape identification. Each of the 54 shapes appeared once in each condition, for 162 identification responses per subject.

In the 2-frame displays, each subsampled dot appears for exactly 2 consecutive new frames. Half of the dots are replaced with another random subsample on each new frame. This introduces 50% scintillation (density control does not require additional subsampling). In the 3-frame displays, each dot appears for exactly 3 consecutive new frames. One-third of the dots are replaced with another random subsample on each new frame, for 33% scintillation. In the ≤30-frame displays, each dot remains visible for all 30 new frames of the display, with exceptions to eliminate the density cues, which introduced 5% scintillation. This is identical to condition (2) of Expt 1.

Display geometry and calibrated intensities

The identification stimuli, subjects, and viewing conditions are identical to those listed in Method Expt 1. Calibrated intensities were identical to those in that experiment.

METHOD: EXPERIMENT 4 (VISIBILITY)

Conditions

Conditions for Expts 4–6 are listed in Table 2. This experiment required subjects to detect the

presence of uniform planar motion in a two-interval forced-choice (2IFC) paradigm (Fig. 7a). The subject indicated which interval contained the moving stimulus. Guessing baserate is 50%. Stimuli consisted either of normal light dots on a gray background (conditions 15–19), or polarity alternating light and dark dots on the background (conditions 20–24). The five conditions of each type are measures of motion-direction at five levels of the "standard" (condition 2, Expt 1) dot intensity (increments or decrements). For MSL, the intensity conditions were 17%, 25%, 33%, 42% and 5[illegible] of standard. For JBL, the intensity conditions were 33%, 50%, 67%, 83% and 100% of standard. The 10 conditions each were tested 20 times per block in random order, for 5 blocks, or a total of 1000 trials per subject.

Display geometry and calibrated intensities

Each interval of the display consisted of a $\frac{1}{3}$ sec fixation spot, $\frac{1}{3}$ sec blank screen, followed by 1.067 sec (16 frames at 15 frames/sec) of stimulus. Non-motion intervals displayed uniform gray fields. Motion intervals displayed a sequence of approximately 17 random dots in a 48 × 48 pixel (0.97 by 1.1 deg) patch (0.0075 dots/pixel, or 16 dots/deg^2 average density) moving left or right by 1 pixel/frame, or approximately 0.35 deg/sec. The viewing conditions were identical to those described above for Experiment 1. For MSL, the background intensity was 31.0 cd/m^2, and the intensity increment or decrement was 13.2 μcd/dot at 100% standard intensity. For JBL, the background was 32.0 cd/m^2 and the increment or decrement was 16.9 μcd/dot at 100% standard intensity.

METHOD: EXPERIMENT 5 (MOTION DIRECTION)

Conditions

The task in this experiment was discrimination of leftward from rightward motion of dots within a square in the center of a larger field (Fig. 8a). The stimuli were a uniform field of dots of approximately the same density as the shape identification stimuli of Expts 1–3. The drift speed of dots in the central square (0.35 deg/sec) was approximately the average of ground dots at the edges of the shape identification stimulus, or approximately one-eighth of the peak velocity in that stimulus. In the 3D shape identification stimuli, peak speed is achieved for only one or two frames and then only at the exact center of a peak or valley. Most dots in the vicinity of a peak or valley have an average speed of one-half peak speed or less. The selection of drift speed for this direction of motion control is considered in the Results section.

The dots were either all white on gray (standard) (conditions 25–29) or alternated in polarity (conditions 30–34) from frame to frame. Standard and alternating images were crossed with five increment intensity levels at 33%, 50%, 67%, 83% and 100% of the standard increment/decrement. Each of the 10 conditions had 200 samples, 100 with each movement direction, for a total of 1000 direction judgements per subject.

Display geometry and calibrated intensities

Each trial consisted of a $\frac{1}{3}$ sec spot, $\frac{1}{3}$ sec blank gray frame, and 1 sec motion display, followed by a blank frame during the response interval. The image was 200 × 200 pixels, 4.1 by 4.6 deg at a viewing distance of 1.6 m. This included a dynamic noise background, with a moving center of 48 × 48 pixels. Dot density was approximately 16 dots/deg^2, and drift velocity was 1 pixel/frame, or approximately 2.3 min arc/frame, or 0.35 deg/sec. The viewing conditions and calibrated *standard* intensities are the same as those in Method Expt 1.

METHOD: EXPERIMENT 6 (MOTION SEGMENTATION)

Conditions

The task in this experiment was motion segmentation. Each display consisted of a 3 × 3 grid of patches of planar motion, with eight patches drifting left (in a left-drifting surround) and one patch drifting right, or vice versa (Fig. 9a). The subject's task was to name the location and direction of the odd motion.

There were two conditions in this experiment: bright dots of standard intensity (35), and dots of alternating polarity (36) on a gray ground.

For JBL, all conditions were intermixed, such that each of three blocks showed 72 stimuli from condition (35), and 54 stimuli from each of condition (36) and a third condition which we do not report here. For MSL, two blocks had 90 trials each of conditions (35) and (36).

Display geometry and calibrated intensities

Each image was 200 × 200 pixels or 4.05 by 4.62 deg at a viewing distance of 1.6 m. Each motion patch was 48 × 48 pixels filled with dots at a density of approximately 17 dots/patch (0.0075 dots/pixel, or ~16 dots/deg²). of a 1 pixel/frame drift. The background moved in the same direction as the common-motion patches; the odd-motion patch moved in the opposite direction. Other viewing conditions were the same as in previous experiments. For MSL, background intensity was 31.0 cd/m², with increment-decrement intensity of 13.2 μcd/dot for conditions (1) and (2). For JBL, background intensity was 38.0 cd/m², with increment/decrement intensity of 19.2 μcd/dot in conditions (1) and (2), and 9.6 μcd/dot for the equated condition (3).

RESULTS

Shape identification

Elimination of the density cue (Experiment 1). When a surface is depicted by a random sampling of surface points which then undergo rotation, local regions of higher or lower dot density change over rotation. To assess the possibility that these changes in dot density *per se* can be used as cues to 3D shape, identification performance for image sequences that include both motion and density cues is compared to those in which density cues are eliminated, or in which only the density (but not the motion cues) are preserved. (See Method Expt 1 for experimental details.) Relevant individual subject data are shown in Fig. 4. (These results were initially reported in Sperling et al., 1989.) Eliminating density cues from motion sequences has only a small effect on the subjects' ability to identify shape from strong structure-from-motion stimuli, which may actually be due to introduction of scintillation. One of the three subjects (MSL) was able to perform significantly above the 1.9% guessing baserate (29.6%) with density cues alone in the absence of motion cues, by using a sophisticated guessing strategy. Since our conditions involve the disruption of strong input to low level motion systems, it was desirable to eliminate any cue, such as density, which might contaminate estimates of shape identification with weak structure from motion image sequences. Therefore, all other displays exclude the density cue. All critical

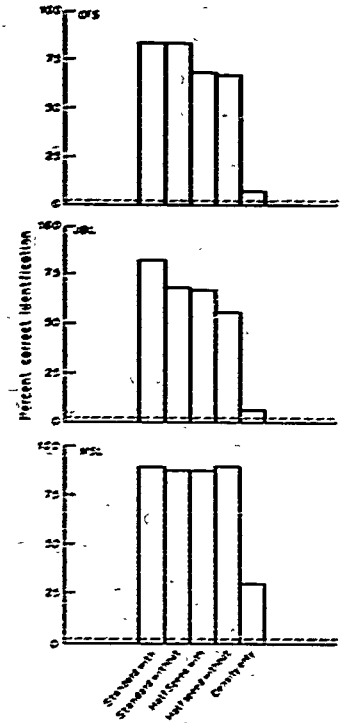


Fig. 4. Shape identification performance for normal displays with and without density cues, and for the density only displays. Performance range is from 0 to 100%, with a guessing baserate of 1.9%. The three panels show data for individual subjects.

image sequences were constructed to have uniform dot density in local regions of the image plane.

Standard sequence: motion without density cue, standard and half-speed (Experiment 1). Percent correct 3D shape identification is shown in Fig. 5. Standard errors of all proportions in the figure are less than 6%; chance is 1.9%. The 3D shape task is illustrated in Fig. 3. *Standard* sequence conditions display sampled dots which are a fixed increment brighter than the gray background. Percent identification levels are shown for "standard" rotation speed (sinusoidal rotation of amplitude 25 deg and period 30 frames, at frame rate of 15 new frames/sec), and for half speed (7.5 new frames/sec). The

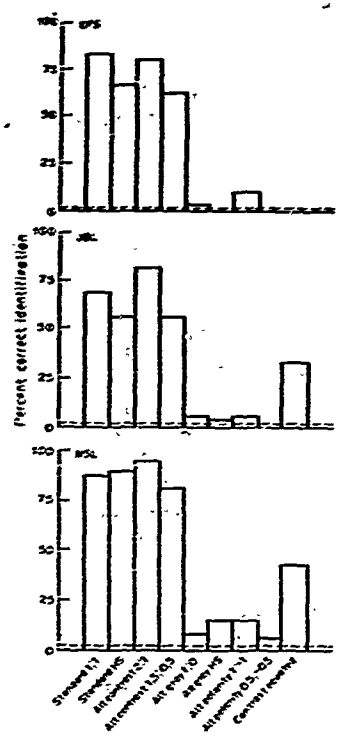


Fig. 5. Shape identification performance for standard displays, alternating gray frame displays, alternating polarity displays and a number of control displays. Performance range is from 0 to 100%, with a guessing baserate of 1.9%. The three panels show data for individual subjects. (Contrast equated condition unavailable for subject CFS)

average percent correct is similar for both speeds, with half-speed slightly less for subjects JBL and CFS.

Gray frame dilution (Experiment 1). By interspersing a background level (gray) blank frame between each frame depicting points of the object, we presented direction-ambiguous information to first-order motion mechanisms while maintaining the visibility of the dot features in any given frame (see Discussion section: Fourier Analysis of the Stimuli). There were two variants of this manipulation: one which equated the viewing time for each new image seen, but consequently slowing the rotation rate of the stimulus in time; and the other which replaced every other new stimulus frame with a blank frame, but equated effective rotation rate. Both of these variants destroyed the ability to recovery shape information from the stimulus (see Fig. 5). Only one of three subjects (MSL) maintained significantly above chance performance (average of 11%) on image sequences with alternating gray frames. Although this represents above chance identification performance, it is dramatically worse than his identification performance of nearly 90% with the unperturbed standard sequence. Rotation speed in these ranges had only small effects on either standard or alternating-gray conditions, and thus can not account for the impact of alternating gray frames on 3D shape performance.

Alternating polarity (Experiment 1). In polarity alternation, the stimulus tokens (subsampled dots on the shape surface) alternate between intensity increments and decrements (light on gray than dark on gray) on each frame. Adjacent image frames primarily support motion signals of the incorrect sign in the first-order system. Analysis of the change in location of these motion signals over many frames, or analysis following some form of rectification (second order, or non-Fourier analysis, see Chubb & Sperling, 1988b) could support the correct motion interpretation. Two levels of polarity alternation were examined, one with light dots equal in intensity to those in standard image sequences and one with light dots half the intensity of those in standard image sequences. In both cases, the dark dots were symmetrically below the background level. Again, disrupting the input to low level motion systems reduced shape identification performance to near guessing baserates. Only one of three subjects (MSL) retained above-chance identification on polarity alternation stimuli (average of 10%).

Intensity alternation stimuli (Experiment 1). Introducing blank (gray) frames between every stimulus frame in an image sequence causes ambiguous signals in the first-order motion systems. Introducing polarity reversal caused direction-reversed signals in the first-order motion systems. Both manipulations also introduce whole-screen flicker, stimulus frames including intensified dots appear every other new frame for a flicker frequency of 7.5 Hz. We included two *contrast alternation* (without polarity alternation) conditions, which also exhibit whole-screen flicker at 7.5 Hz, both of which

sustain performance levels close to that of the standard stimulus.

One flicker control alternated the intensity of stimulus points between the intensity level in normal displays and twice that. This stimulus is the sum of the standard stimulus and the gray frame stimulus. The other flicker control alternated between 1.5 and 0.5 the standard levels. This stimulus is the sum of a half contrast standard and the full-contrast gray frame stimulus. Alternatively, this stimulus can be decomposed into a standard stimulus plus a half-contrast polarity alternation stimulus (i.e. a high-flicker added stimulus). The performance levels on both control conditions are quite consistent with a Fourier power (first-order) analysis of these sequences (see the Discussion). Thus, addition of flicker *per se* does not account for the decrements in performance for alternating-gray and alternating-polarity displays.

Equated intensity control (Experiment 2). We have demonstrated that gray frame alternation and polarity alternation both severely disrupt the ability of subjects to extract 3D shape from an image sequence which allows highly accurate 3D shape identification under standard display conditions. However, perhaps this disruption is not unique to the recovery of depth information. Perhaps it simply reflects a general disruption in visibility or motion discrimination. In order to control for this possibility, we constructed equated-intensity controls based on performance in simple direction-of-motion discrimination. The details of the direction discrimination data are described below and in the Method for Expt 5. By reducing the intensity (lowering contrast and hence visibility) of a standard (light on gray background) planar motion stimulus, it is possible to make it equivalent to a full-intensity polarity alternation stimulus for the purposes of left-right direction-discrimination. The direction-discrimination displays present a patch of moving dots of approximately the same area as a bump in the 3D shape displays. Having found the equivalent reduced-contrast standard stimulus, we then compared 3D shape discrimination for the two stimuli (reduced-contrast normal, full-contrast polarity alternation). These results are shown on the extreme right in Fig. 5 for MSL and JBL. If the effect of polarity alternation can be attributed solely to a visibility-related decrement, then the equivalent intensity condition should have yielded equal shape identification performance to that for polarity alternation. In fact, lowering intensity adversely affected shape identification, but levels were still well above those for shape identification from polarity alternation displays. The percent identification for standard, equivalent intensity and polarity alternation conditions were 87%, 43% and 15%, respectively, for MSL, and 69%, 33% and 6%, respectively, for JBL. (Standard error of the 43% and 33% equated contrast conditions is ±6%.)

Tracking disruption—lifetimes (Experiment 3). We have shown that conditions which disrupt input to low-level motion analyzers also eliminate the ability to perceive three-dimensional shape, at least in the conditions of our experiments. It is interesting to contrast this with a manipulation which eliminates the ability to track individual image features (dots) over multiple frames. Models that emphasize the extraction of specific image features and their image plane location (Hoffman & Bennett, 1985; Ullman, 1979, 1985, etc.) might predict that eliminating feature stability should have an equally large impact on the shape identification. We investigated this hypothesis by comparing feature stability over a full 30 frame image sequence with stimuli in which features (surface dots) were stable for only 3 and 2 frames, after which they were replaced with a different random sample of dots (Fig. 6a). The shape identification data are shown in Fig. 6b. For two subjects (MSL, CFS), reducing tracking to two frames (and increasing scintillation substantially) had very little effect on performance. A third subject's (JBL) two-frame lifetime identification performance was about 54% of normal. While this was a 2 × loss, it was a much smaller loss than the 10 × loss induced by polarity alternation for JBL. Thus, feature-tracking models of the kinetic depth effect appear unable to account for the performance in our experiments.

Motion visibility, discrimination and segmentation

This section compares the disruptive effects of polarity alternation on 3D structure-from-motion (shape identification) to its effects on visibility, direction-of-motion discrimination and segmentation.

Motion visibility (Experiment 4). Subjects were asked to detect which of two temporal intervals contained a motion stimulus and which contained a uniform field of background intensity. The motion stimulus was either a

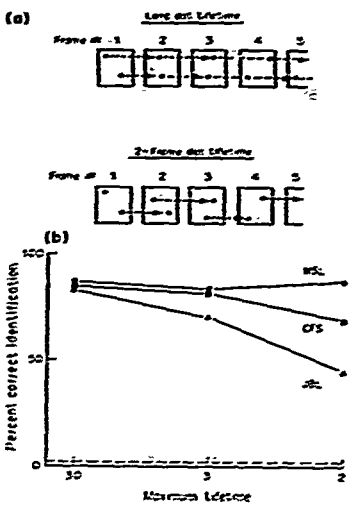


Fig. 6. (a) Illustration of the construction of two-frame lifetime displays, as well as standard construction. In the top panel, sampled dots remain visible in all frames of the display. In the standard 30-frame condition, control for density does actually introduced 5% scintillation, for an expected lifetime of 20 frames. In the bottom panel, sampled dots remain only for two frames, and are then replaced by another sample. (b) Percent shape identification for three subjects in each of the lifetime display conditions. Guessing baserate is 1.9%. Shape identification is little affected by the lifetime manipulation. The small decline may be a consequence of scintillation not loss of trajectory information.

random dot field moving at uniform velocity to the right or left, or a polarity alternation version of the same stimulus. The display is schematically illustrated in Fig. 7a. The size of the region was approximately that of a single peak or valley in the shape displays, of approximately the same dot density and a representative velocity (between that of the ground and maximal velocity of a peak or valley). (See Method Expt 4 for details.) Detection may reflect contributions by nonmotion systems. For example, Watson and Ahumada (1985) claim that detection of moving stimuli with velocity less than 2 deg/sec is performed by non-motion systems.

The detection data are shown in Fig. 7b. Across a range of stimulus intensity increments (17%–50% of standard level intensity for MSL, 33%–100% of standard level intensity for JBL), the effect of polarity alternation was small. For MSL, standard and polarity alternation displays (averaged across contrasts) yielded 73% and 74% correct detection, respectively. For JBL, the figures were 83% and 90%, respectively. Whereas polarity alternation almost destroys the ability to extract three-dimensional shape, it may slightly improve stimulus detection relative to standard displays for our conditions. Detection accuracy with polarity alternation is essentially perfect at intensity levels comparable to those used in the 3D shape experiment (MSL at 50% intensity is 95% correct, and JBL at 100% intensity is 96% correct).

The small effect of polarity alternation on detection performance is consistent with the near-symmetry of increment and decrement

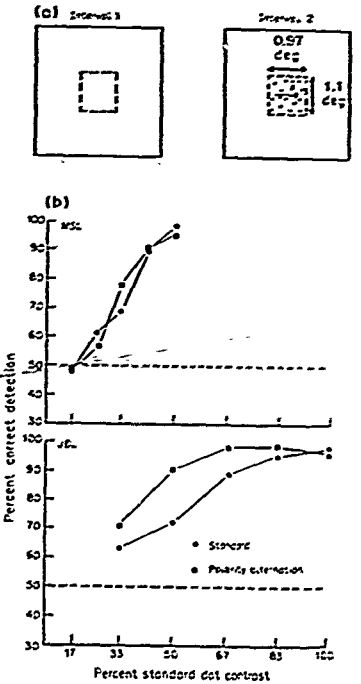


Fig. 7. (a) Illustration of the two-interval forced choice (2IFC) paradigm for the motion visibility task. Subjects judged which 1 sec interval contained a stimulus, and which interval was blank. (b) Percent detection of a planar motion display in the 2IFC task. Detection is measured for standard and polarity alternation image sequences as a function of dot intensity (expressed as a percentage of a standard intensity). Guessing baserate is 50%.

thresholds in small-target pedestal detection experiments (Krauskopf, 1980; Rashbass, 1970; Rouse, 1974), although some studies find decrements slightly easier to detect (Patel & Jones, 1968; Short, 1966). Alternatively, the fundamental flicker component of the polarity alternation stimulus is 7.5 Hz, approximately at the peak of the flicker sensitivity function (Watson, 1986), which suggests that polarity alternation may be most sensitively detected by flicker-sensitive mechanisms.

Direction-of-motion discrimination (Experiment 5). Subjects were asked to discriminate the direction of motion (right or left) of a small patch of random dots moving with uniform velocity (Fig. 8a). The dots were either always light against the background, or alternated polarity from frame to frame. Discrimination was examined over a range of intensity increments (or decrements) per dot. (See Method Expt 5 for details.)

Direction discrimination data are shown for two subjects in Fig. 8b. Polarity alternation impaired subjects' ability to discriminate motion direction: averaged over intensity level, standard and polarity alternation conditions yielded 85% and 69% correct, respectively, for subject MSL, and 90% and 67% respectively for JBL. However, at the intensity levels that were investigated in the shape identification experiments, levels of direction discrimination for polarity alternation stimuli were good: 87% correct for MSL and 88% correct for JBL. Intensity-based decrements for standard displays in this experiment were used to select the "equated intensity" condition listed above for shape identification.

The patch size in the direction-of-motion displays were selected to be approximately the size of a bump or depression in the shape displays. The speed of drift (0.35 deg/sec) was selected to be representative of the modest speeds in many points of the 3D shape displays, where peak speeds may range up to 2.5 deg sec. Based on data from direction of motion discrimination in near-threshold sine wave stimuli (Ball & Sekuler, 1979; Burr & Ross, 1982; Green, 1983; Watson, Thompson, Murphy, & Nachmias, 1980) and theoretical computations on direction of motion discrimination for random dot stimuli (Nakayama, 1985; van Doorn & Koenderink, 1982), we picked the weakest motion stimulus that could be derived from the 3D shape task: the slowest reasonable speed and approximately the same number of dots in the displays to be comparable. That the direction of

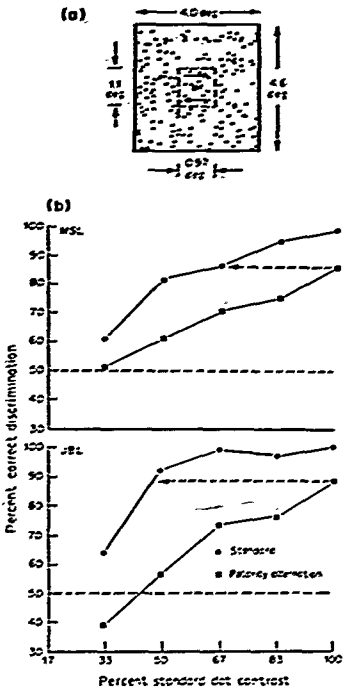


Fig. 8 (a) Schematic illustration of the motion direction discrimination task. Outer dots were dynamic noise, dots in the central patch drifted left or right at 0.35 deg sec. Subjects judged the direction of motion of dots in the central patch. (b) Percent correct discrimination of the direction of motion in the 2D motion-direction display. Discrimination is shown as a function of the intensity increment (as a percent of the "standard" intensity increment), of the stimulus dots on a gray background. The intensity increment where the dashed line-and-arrow intersects the performance line for standard displays equates standard (at reduced intensities) and polarity alternation displays (at standard intensities). The guessing baserate is 50%. Panels show the data of different subjects.

motion of this stimulus is nearly always judged correctly at standard intensities implies that direction of motion at a *single* location is almost completely intact when 3D shape identification is at zero.

In two-frame experiments or multi-frame experiments where two frames appear alternately, polarity alternation may lead to below chance performance on direction discrimination (Anstis, 1970). Polarity alternation excites first-order (Fourier) spatio-temporal sensors for

motion opposite to the veridical direction, as schematically illustrated in Fig. 2e. Here, in multi-frame movement, the (temporally and spatially) local support for movement in the opposite direction is apparently more than offset by second-order (nonFourier) processes sufficiently often that direction discrimination rarely falls below 50%. Chubb and Sperling (1988a, 1989a, b) show that the relative dominance of the first-order and second-order information in polarity alternation stimuli depends on the spatial scale (near viewing distances favor second-order information).

Motion segmentation (Experiment 6). In contrast with simple detection or discrimination of motion direction, a more complex direction task did show decrements in performance more comparable to those seen in shape identification. We developed a motion segregation paradigm in which nine small patches of uniformly moving dots were presented as a 3 × 3 grid embedded in a border of moving random dots (Fig. 9a). All but one patch depicted motion in the same direction (left or right), while the odd patch depicted motion in the opposite direction. The stimulus dots either remained above the background level (light on gray), or alternated polarity. (See Method Expt 6 for details.) In this situation, polarity alternation had a large impact on selection of the odd patch. MSL reported 95% correct locations with the standard display, but only 22.2% with polarity alternation. JBL reported 84% correct and 10.5% respectively (chance = 11.1%) (Fig. 9b). The accuracy levels for polarity alternation displays are consistent with sophisticated guessing (see Discussion).

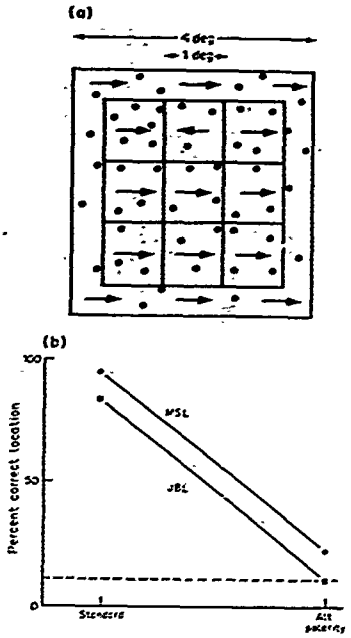


Fig. 9. (a) Schematic illustration of the nine-location forced-choice (9LFC) motion segmentation display. Subjects judged the location of the single patch moving opposite in direction to the other eight. (b) Percent correct location judgement for the 9LFC task for standard and alternating polarity displays the two subjects. Guessing baseline is 11.1% (1 in 9).

DISCUSSION

Fourier and nonFourier inputs to structure from motion

Vivid 3D shape percepts which allow accurate 3D shape identification can arise from appropriately constructed 2D image sequences depicting projections of those shapes under rotational motion. Typically these 2D sequences provide good input to first-order spatio-temporal ("Fourier") motion analyzers. In order to determine whether strong Fourier motion is a prerequisite to shape extraction, we examined display manipulations which maintain the identity-correspondence between points in successive frames, but disrupt first-order analysis. Interleaving blank frames or alternating token contrast-polarity both had devastating consequences for the ability to identify 3D shape in our displays. The inability to recover shape was not due to overall display flicker since same-sign alteration in the intensity levels of particular tokens did not seriously disrupt performance. Subjectively, a sensation of local motion was maintained, and selected points could still be tracked. Nonetheless, this information was not adequate to support shape identification.

The dependence of 3D shape perception on unambiguous first-order (Fourier) motion inputs suggests that, for our stimuli, direction and velocity serve as the primary input to a subsequent shape-extraction (structure from motion computation, e.g., Koenderink & van Doorn, 1986). Obviously the velocity information must be computed simultaneously or nearly

simultaneously at several locations in order to perform the 3D shape task.

The main alternatives to local velocity-based computations depend on geometric analyses of identified feature elements and operate over more than two frames (e.g. Ullman, 1986). These alternative schemes are challenged by our finding that shape extraction is little affected by change in feature elements as often as every two frames. Further, subsequent work (Landy, Dosher, Sperling & Perkins, 1988) shows that motion displays of only two-frames also support moderately good shape identification.

Williams and Phillips (1986, 1987) report what they consider a surprising perceptual phenomenon of perceiving a 3D shape in a random-dot flow field. We interpret their finding here as further evidence that a local velocity computation is the basis of perception of 3D shape. In their dynamic 2D displays, dots execute a random walk of constant step size, with displacement angle chosen from a uniform distribution with a range less than 150 deg. Subjects perceive a rotating and translating 3D cylinder. In these stochastic displays, velocity information is very similar to the local velocity information in a cylinder with dots sprinkled through its volume, rotating rigidly and translating along its axis of rotation (e.g. as displayed by Dosher, Landy & Sperling, 1989).* As in our experiments, the momentary distribution of velocities, not the stochastic trajectories of individual dots, determines the 3D percept.

3D shape extraction is especially impaired in displays that have contradictory or ambiguous first-order (Fourier) information. Control experiments demonstrated that contrast-polarity alternation, which essentially eliminated 3D shape identification, nonetheless left the detection judgement and the direction-of-motion judgement for a small isolated moving patch quite high. Motion segmentation, which requires analysis of motion direction in a number of local display regions, was also profoundly affected by polarity alternation.

A Fourier computation for the strength of first-order motion perception

Up to this point, we have talked in generalities about Fourier and non-Fourier computations of motion direction. Here we propose some very simple, specific, Fourier computations that account quite well for the results that we have attributed to first-order motion processes. The computation proceeds as follows.

(1) Compute the Fourier transform of the stimulus as it was viewed by the observer, i.e., with the correct visual angle and an accurate description of the display that was actually produced. Compute the power $p(\omega_x,\omega_t)$ of each spatio-temporal frequency component.
(2) Retain only the power p_t that exceeds a small threshold $\varepsilon > 0$, i.e. $p_t(\omega_x,\omega_t) = \max[p(\omega_x,\omega_t) - \varepsilon, 0]$.
(3) Retain only the Fourier components that fall within a window of visibility (Watson, Ahumada & Farrell, 1986) that includes all spatial frequencies greater than zero and less than or equal to 30 cycles per degree of visual angle and all temporal frequencies greater than zero and less than or equal to 30 Hz, viz. $(0 < |\omega_x|, |\omega_t| \leq 30)$.
(4) The *net directional power*, *DP*, of all frequencies within the window of visibility is the rightward power minus the leftward power:

$$DP = \sum_{\substack{\omega_x\,\omega_t<0 \\ |\omega_x|,|\omega_t|\leq 30}} p_t(\omega_x,\omega_t) - \sum_{\substack{\omega_x\,\omega_t>0 \\ |\omega_x|,|\omega_t|\leq 30}} p_t(\omega_x,\omega_t).$$

The computation gives equal weight to all motion components within the window of visibility and zero weight to all components outside the window. In a more refined analysis, it might be useful to weight spatial frequencies according to a contrast sensitivity function. However, it is not obvious how to weight signals that are above threshold. For practical purposes, it turns out that the exact size of the window of visibility has little influence on relative *DP*s for the stimuli considered here.

Basically, the left-minus-right-difference, summed over all frequencies, is similar to the computation that is carried out by previously proposed first-order motion models. For example, within its window, an elaborated Reichardt motion detector (van Santen & Sperling, 1984)

*In the display of a transparent cylinder filled with dots, rotating around a central vertical axis and translating upward, dots viewed through the middle of the cylinder have a greater range of lateral motion velocities and dots at the 2D edges have a smaller range of velocities; in Williams and Phillips' random flow field, there is a wide range of velocities throughout the display. However, at the edges, dots disappear and re-appear, this scintillation (as in Expt 2) reduces the magnitude of perceived depth; mean lateral velocity in both areas is zero. The effective flow fields for these differently constructed stimuli actually are quite similar.

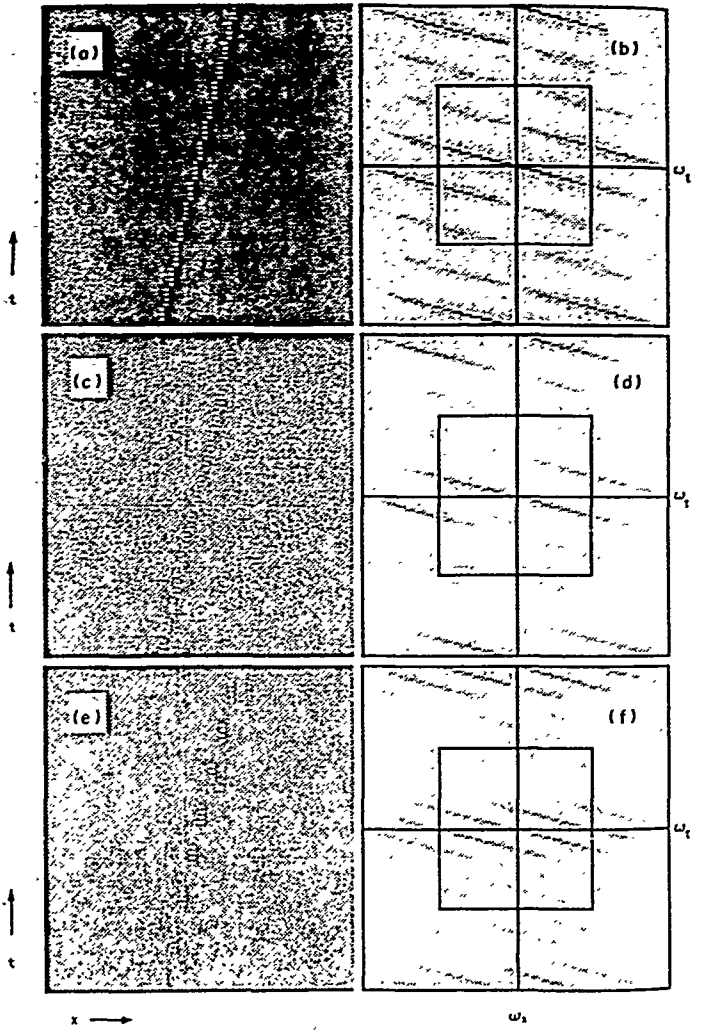


Fig 10 (a-f)

Fig 10 Stimulus representations and corresponding Fourier energy spectra typical of various displays moving at a rate of 0.35 deg sec. The abscissa is (horizontal) spatial location, and the ordinate is time resolution of 60 Hz. The stimulus is either light or dark increments or decrements on a gray background inner boxes represent the window of visibility, assumed to resolve less than or equal to 30 c/deg and less consistent with the intended direction of motion. The upper right (or lower left) quadrant of the spectra spectrum for the "standard" stimulus are shown in (a, b), for the half-contrast standard stimulus in (c d) contrast 2:1 in (i, j), and for the

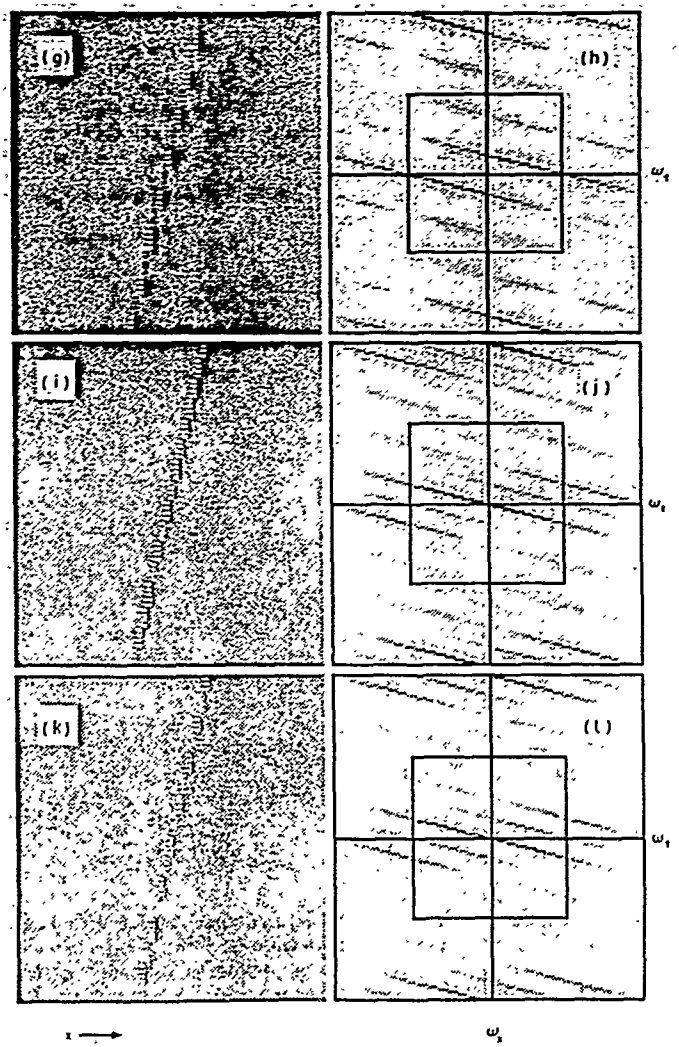


Fig. 10 (g-l)

Fig. 10 (*Continued*)

conditions (a, c, e, g, i, k) Each stimulus representation depicts 1 07 sec of planar motion of a single dot The representation assumes spatial resolution of 60 cycles per degree of visual angle; and temporal (b, d, f, h, j, l) The corresponding Fourier spectra are shown on ω_x (abscissa), ω_t (ordinate) axes. The than or equal to 30 Hz The upper left (or lower right) quadrant of the spectra represent power at (ω_x, ω_t) represent power at (ω_x, ω_t) consistent with the unintended direction of motion The representation and for the alternating-gray stimulus in (e, f), for the alternating-polarity stimulus in (g, h), for the alternating alternating contrast 1 5 0 5 in (k, l)

computes the algebraic sum of all velocity inputs that differ in temporal frequency. Velocity inputs that have the same temporal frequency (and therefore differ only in spatial frequency) are processed by detectors of different scales, sensitive to different spatial frequencies. Outputs of different detectors are combined at the next higher level (e.g. Adelson & Bergen, 1986).

A real detector, localized in space and time, cannot have the perfect resolution of a Fourier analysis of the entire x,y,t stimulus. The entire Fourier analysis is most appropriate for analyzing local areas where movement can be regarded as uniform and homogeneous. Even with all these qualifications, the straightforward Fourier analysis of the dot movement-patterns is quite informative.

Fourier analysis of the stimuli

The space-time (x,t) representations of a single dot element in each of the motion stimuli for our main conditions is shown in the left hand panels of Fig. 10. The Fourier power spectra for those stimuli are shown in the right hand panels of Fig. 10. Figure 10a represents a dot moving from left to right over frames. The dot is the standard intensity on the neutral background. The abscissa represents 1.07 deg of spatial position x from left to right; the ordinate represents a 1.07 sec interval of time, t, from bottom to top. The representation assumes a sampling density of 120 samples per degree of visual angle and 120 samples per second to yield temporal discrimination up to 60 Hz and spatial discrimination up to 60 c/deg of visual angle. (In this representation, the four refreshes of each new image frame are seen as four repeats at the same location in alternate 1/120 sec samples. The illuminated dots on our display are depicted as 2 adjacent spatial samples.) The steep space-time function reflects the fact that our stimuli move relatively slowly (0.35 deg/sec). Figure 10b shows the corresponding Fourier power spectrum. The abscissa is ω_x and the ordinate is ω_t; the axes cross at $\omega_x = \omega_t = 0$.

If the standard motion stimulus were moving continuously in space and time, essentially all of its components would be at the intended direction and speed. Because it is sampled in time (60 Hz refresh and 15 new frames/sec) and in space (by the resolution of the pixel array) it contains ambiguous temporal and spatial components. Most of the power is in the intended direction and velocity (upper left and, symmetrically, lower right quadrants). But there is a surprising amount of power in the unintended direction as well (upper right, and symmetrically, lower left quadrants). The $(0 < |\omega_x|,|\omega_t| \leq 30)$ window of visibility is shown as the inner square in Fig. 10. The computed DP strongly favors the intended direction by 5:1. Figures 10c and d show the stimulus representation and Fourier energy spectrum of a standard stimulus at half intensity (approximately that of the contrast-equated control). The transform is the same as Fig. 10b, but of half power. With $\varepsilon = 0$, the computed DP is exactly half; with $\varepsilon > 0$, the computed DP is less than half.

Figures 10e and f show the stimulus representation and spectrum for the alternating gray frame stimulus. In the case of gray-frame stimuli, power at the intended direction and velocity is halved, and approximately balanced by power dispersed over a range of velocities in the opposite direction.

Figures 10g and h show the stimulus representation and spectrum for the alternating-contrast polarity stimulus. In this case, the net directional power DP is of very slightly lower magnitude than for the standard stimulus, but favors the unintended over the intended direction (more power in the upper right and lower left quadrants).

Figures 10i and j show the stimulus with contrast alternation between 2× and 1× the standard intensity. This stimulus can be viewed as the sum of the standard stimulus and the alternating-gray stimulus. Although the 2:1 contrast-alternating stimulus has some of the diffuse power of the alternating-gray stimulus, 2:1 contrast alternation puts more power into the intended direction and velocity than even the standard stimulus. Figures 10k and 10l are for stimuli with contrast alternation between 1.5× and 0.5× the standard intensity. This 1.5:0.5 contrast-alternating stimulus can be viewed as the sum of the half-intensity standard stimulus and the alternating-gray stimulus. The computed DP is slightly lower than for the standard stimulus.

Tasks

The kinds of information needed for good performance in the various tasks is summarized in Fig. 11 and, along with the relation to computed DP, is explained below.

Detection. In Expt 4, we noted that simple two-interval forced choice detection (2IFC Detection) of a single local patch of moving dots

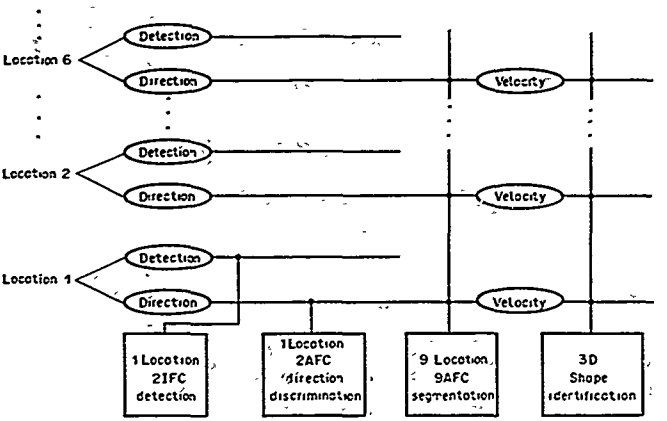


Fig. 11. A schematic illustration of the kinds of information required in order to perform each of the experimental tasks. The simple 2IFC detection task may reflect the output of non-motion systems in a single location. The 2AFC discrimination of motion direction task requires the output of a motion direction mechanism in a single location. The 9LFC motion segmentation task requires the output of motion direction mechanisms in a number of locations nearly simultaneously. The 3D shape task requires direction and speed information from a number of locations nearly simultaneously.

is probably accomplished by other systems than the motion systems. The equality (or near equality) of detection with standard and polarity alternation displays insures that polarity alternation did not result in peripheral cancellation of the input stimulus.

Direction. Discrimination between left and right motion direction (two-alternative forced choice, 2AFC Direction) minimally requires direction (but not necessarily velocity) analysis by a motion detection system in a single location (Fig. 11). As shown by the Fourier spectrum of Fig. 10h, a first-order analysis of a polarity-alternation stimulus would support the unintended (opposite) direction of movement. A second-order analysis based on full-wave rectification would yield the correct direction and velocity. In full-wave rectification, the sign of contrast is lost, and the standard stimulus would be recovered. 2AFC-direction performance is impaired by polarity alternation, but still well above chance for a wide range of contrasts. Polarity alternation leads to high levels (about 88% correct) of 2AFC-direction performance at "standard" contrasts, hence, perceptual second-order analysis occurs under these conditions. But, alternating-contrast polarity stimuli require higher contrasts to yield equal direction-discrimination than do standard stimuli which stimulate first plus second-order systems. This might reflect power loss in the second-order analysis, the need to overcome conflicting first-order information, or both.

Motion segmentation In order to isolate which of 9 patches is moving in a direction opposite to the others requires that direction of motion be assessed in several locations (Fig. 11) We examine the consequences of observing (correctly perceiving the direction of motion in) n of the 9 locations. Observing just one patch, which is sufficient for the 2AFC-Direction task would lead to chance performance of one-in-nine locations—identical to the guessing level without seeing the display. Observing any two patches could improve performance by sophisticated guessing. That is, if the two patches move oppositely, then one of them is the target, if they move in the same direction, one of the remaining 7 is the target The probability of sampling two opposite direction locations times a guessing accuracy of 1/2 plus the probability of sampling two same directions times a guessing accuracy of 1/7 yields an estimate of 22 2% correct. Observing any three or more patches could improve performance by a combination of informed judgements and sophisticated guessing, etc The data for polarity alternation do not require us to consider more than two

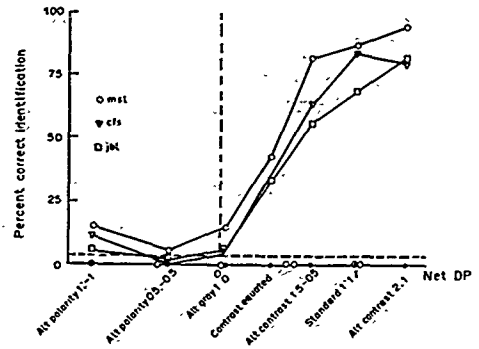


Fig 12. The relation between 3D shape identification performance and computed *net directional power DP* within the window of visibility and above a threshold ε. Solid circles on the abscissa are values of *DP* computed from the spectra in Fig. 10, panels (b), (d), etc., for an ε of 0.12× the maximum power value in the spectrum of the standard stimulus. Open circles on the abscissa are the values of *DP* computed for an ε of 0. (The rank order of conditions under the two computations is the same.) The 3D shape identification performance is monotone with *DP* for all reasonable values of $\varepsilon \geq 0$.

observations. Performance for polarity alternating stimuli in the odd-in-nine motion segmentation task was indistinguishable from the simple 1 in 9 baseline (11%) for one subject (10%), and slightly above the 1 in 9 baseline for another (22%), which could be achieved by sampling only two locations.

Motion segregation, like shape extraction, may be dependent on strong Fourier input largely because it requires evaluation of motion signals at more than one location nearly simultaneously. The second-order motion system operates primarily foveally (Chubb & Sperling, 1988b). Two locations might be successively fixated in our 1 sec displays. For standard displays, performance in this task is excellent (85-95%). By similar computations, this would require observation of approximately 7 locations. Thus, first-order information supports direction of motion analysis at a number of directions simultaneously, while second-order information can support direction of motion analysis at only one or two.

3D shape. The simplest solution to the 3D shape identification task requires simultaneous, or nearly simultaneous, knowledge of the motion-direction information (and possibly also the velocity) at the six bump locations (Sperling et al., 1989). The principle is that, to a first and adequate approximation, dots on bumps move in one direction, dots in depressions move in the opposite direction, and dots on the ground plane move very little. Thus, to solve the 3D-shape task, motion has to be categorized into 3 categories (leftward, rightward, and near zero) at a number of locations simultaneously * Although the 3D-shape identification task could, in principle, be carried out with only this very coarse velocity information, more information usually is used. For example, in a version of the 3D-shape identification task with different bump heights, subjects can quickly discriminate

*At certain moments during the rotation, dots on bumps move opposite to ground dots, and at other moments dots on depressions move opposite to ground dots. To solve the task by motion direction only would require sampling at least three frames. That is, to observe any motion at all, requires two frames. Since there are only two categories of motion-direction response, from the motion observed in the first two frames, only two categories of dots could be observed (e.g. left or rightward moving). By observing a third frame, some of the dots that were categorized together in the first two frames could be differentiated (e.g. initially leftward, then rightward) and this could be used, in principle, to set up the three categories of dots (forward, center, behind) needed to solve the 3D shape discrimination task. However, we show (Landy et al., 1988) that two frames suffice for accurate performance. This means that at least three (moving leftward, moving rightward, not moving) and probably more categories of velocity information are available. Therefore, for the present discussion, we can assume that our 3D shape identification task has access to three-category velocity information. this velocity information obtained simultaneously from (at least) six locations would suffice to solve the task.

three levels of bump height (Sperling et al., 1989). The bump-height discrimination is based on speed.*

Although a sophisticated local velocity computation probably underlies the 3D shape *percept*, for our set of stimuli, the simple (Fourier) net directional power, *DP*, computation offers an adequate account of performance in the 3D shape identification *task*. We assume that net directional power *DP* serves as a measure of the quality of first-order direction information in the various displays. If the 3D shape identification performance with our displays primarily depended on good first-order information, then the performance level for the various displays would increase monotonically with the quality of first-order information—here indexed by *DP*. Figure 12 shows the percent correct identification in the 3D shape task as a function of computed *DP* for the representative 2D motion display (Fig. 10a–l). *DP* is in units of power normalized to the standard stimulus. Identification levels increase monotonically with DP, as expected.

Full-wave rectification of polarity alternation displays (second-order processing) would allow recovery of intended motion signals. However, 3D shape identification performance on these displays is approximately at chance levels (left half of Fig. 12). In principle, systematic *DP* favoring the unintended direction might be used in sophisticated guessing, but apparently is not. Performance on displays with polarity alternation may also reflect conflict between first-order and second-order motion information.

The effect of the power threshold ε in the computation of *DP* may be understood by comparing 3D shape performance in the contrast equated (approximately half-power standard) and 1.5:0.5 contrast-alternation stimuli. Without the power threshold ε entering into computed *DP*, the contrast alternation 1.5:0.5 computed *DP* is only slightly higher than that for the half-intensity standard, while identification levels are quite different. However, even with $\varepsilon = 0$, identification performance is monotone with *DP*. (DP computations with $\varepsilon > 0$ and with $\varepsilon = 0$ are shown, as filled and open circles, respectively, on the abscissa of Fig. 12.) Hence, the 3D shape data are consistent with a *DP* analysis of the outputs from a first-order (Fourier) motion system.

Why first-order motion for 3D shape perception?

First-order (Fourier) motion systems are assumed to be implemented with detectors like those schematized in Fig. 1. Second-order (non-Fourier) motion systems may implement some form of nonlinear transformation on the image intensities prior to further spatio-temporal analysis (see Chubb & Sperling, 1987). The two tasks in which second-order information could not be efficiently utilized, 3D shape recovery and motion segmentation, require information about motion direction (and velocity) in several local regions simultaneously. Hence, our evidence agrees with the evidence of Chubb and Sperling (1988a, b, 1989a, b) that the non-Fourier motion systems are most effective at large spatial scales, with foveal presentation, and do not function well in noncentral locations. For our stimuli, 3D structure was extracted primarily from first-order motion information.

Our stimuli were modestly complex but continuous surfaces in depth. The surfaces were depicted by randomly scattered and unconnected dots. Object transparency (where a portion of the stimulus which is behind a nearer portion of the surface can be seen) was allowed, but rarely occurred. (This form of representation is most similar to defining shape by local texture elements in naturalistic displays.) Precisely what the boundary conditions are on these findings remains to be determined. Because our dot stimuli are small, sparse, and hence of low total contrast power, they may be particularly poor stimuli for a second-order motion system. Prazdny (1986) reported an example of 3D shape from second-order motion stimuli (which do not effectively stimulate first-order mechanisms) for very simple (4 bend) wide wire figures. The wires were depicted by dense random dynamic noise against a background of dense static noise. His shapes were very simple, nonsurface shapes, and were not edited to exclude 2D information about identity. However, his thick wires are a better stimulus (than our dots) for a second-order system due to the large spatial scale.

In a subsequent paper (Landy, Sperling, Dosher & Perkins, 1988), we examine kinetic depth stimuli that are statistically invisible to Fourier detectors. We use various different stimulus tokens (dots, disks, wires) and backgrounds

*To prove that the relevant cue for discriminating bump heights is speed, possible alternative cues, such as distance traversed and the configuration at the point of rotation reversal must be irrelevantly varied so that they can not become artifactual cues

(gray, static random noise), as well as polarity alternation of standard stimuli. For large-scale tokens, polarity alternation is very damaging, but some residual above-chance 3D shape identification appears to be possible. That investigation also supports and generalizes the conclusion that the *primary* substrate of shape identification is strong first-order motion information for stimuli which require analysis of motion in a number of regions simultaneously. However, appropriately constructed displays, which provide a high power stimulus to the second-order motion systems, may support reduced, but above-chance 3D shape analysis.

Acknowledgements—This work was supported by Office of Naval Research, Grant N00014-85-K-007 and by AFOSR, Life Science Directorate, Visual Information Processing Program, Grants No. AFOSR 85-0364 and 88-0140.

REFERENCES

Adelson, E. H. & Bergen, J. R. (1985). Spatiotemperal energy models for the perception of motion. *Journal of the Optical Society of America A, 2,* 284–299.

Adelson, E. H. & Bergen, J. R. (1986). The extraction of spatio-temporal energy in human and machine vision. *Proceedings of Workshop on Motion: Representation and Analysis.* IEEE Computer Society # 696, 151–155.

Andersen, G. J. & Braunstein, M. L. (1983). Dynamic occlusion in the perception of rotation in depth. *Perception and Psychophysics, 34,* 356–362.

Anstis, S. M. (1970). Phi movement as a subtraction process. *Vision Research, 15,* 957–961.

Anstis, S. M. & Rogers, B. J. (1975). Illusory reversal of visual depth and movement during changes of contrast. *Vision Research, 15,* 957–961.

Ball, K. & Sekuler R. (1979). Masking of motion by broad band and filtered directional noise. *Perception and Psychophysics, 26,* 206–214.

Braddick, O. (1973). The masking of apparent motion in random-dot patterns. *Vision Research, 13,* 355–369.

Braddick, O. (1974). A short range process in apparent motion. *Vision Research, 14,* 519–527.

Braunstein, M. L. (1962) Depth perception in rotating dot patterns: Effects of numerosity and perspective. *Journal of Experimental Psychology, 64,* 415–420.

Braunstein, M. L., Hoffman, D. D., Shapiro, L. R., Andersen, G. J. & Bennett, B. M. (1987). Minimum points and views for the recovery of three-dimensional structure. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 335–343.

Burr, D. C. & Ross, J. (1982). Contrast sensitivity at high velocities. *Vision Research, 22,* 479–484.

Burt, P. and Sperling, G. (1981). Time, distance, and feature trade-offs in visual apparent motion. *Psychological Review, 88,* 171–195.

Chubb, C. & Sperling, G. (1987). Drift-balanced random stimuli: A general basis for studying nonFourier motion perception. *Investigative Ophthalmology and Visual Science (Supplement), 28,* 233.

Chubb, C. & Sperling, G. (1988a). Processing stages in non-Fourier motion perception. *Investigative Ophthalmology and Visual Science (Supplement), 29,* 266.

Chubb, C. & Sperling, G. (1988b). Drift-balanced random stimuli: A general basis for studying non-Fourier motion perception. *Journal of the Optical Society of America A: Optics and Image Science, 5,* 1986–2006.

Chubb, C. & Sperling, G. (1989a). Second-order motion perception: Space-time separable mechanisms. *Proceedings: 1989 IEEE Workshop on Motion.* Washington, D.C.: IEEE Computer Society Press, in press.

Chubb, C. & Sperling, G. (1989b). Two motion perception mechanisms revealed by distance driven reversal of apparent motion. *Proceedings of the National Academy of Sciences, U.S.A., 86,* in press.

Clocksin, W. F. (1980). Perception of surface slant and edge labels from optical flow: A computational approach. *Perception, 9,* 253–269.

van Doorn, A. J. & Koenderink, J. J. (1982). Spatial properties of the visual detectability of moving spatial white noise. *Experimental Brain Research, 45,* 189–195.

Dosher, B. A., Landy, M. S. & Sperling, G. (1988). The kinetic depth effect and optic flow: I. 3D Shape from Fourier motion. *Mathematical Studies in Perception and Cognition, 88-4,* NYU Report Series.

Dosher, B. A., Landy, M. S. & Sperling, G. (1989). Ratings of kinetic depth in multi-dot displays. *Journal of Experimental Psychology: Human Perception and Performance,* in press.

Fennema, C. L. & Thompson, W. B. (1979). Velocity determination in scenes containing several moving images. *Computer Graphics and Image Processing, 9,* 301–315.

Foster, D. H. (1969). The response of the human visual system to moving spatially-periodic patterns. *Vision Research, 9,* 577–590.

Foster, D. H. (1971). The response of the human visual system to moving spatially-periodic patterns: Further analysis. *Vision Research, 11,* 57–81.

Green, B. F. (1961). Figure coherence in the kinetic depth effect. *Journal of Experimental Psychology, 62,* 272–282.

Green, M. (1983). Contrast detection and direction discrimination of drifting gratings. *Vision Research, 23,* 281–289.

Harris, M. G. (1986). The perception of moving stimuli: A model of spatiotemporal coding in human vision. *Vision Research, 26,* 1281–1287.

Heeger, D. J. (1987). A model for the extraction of image flow. *Journal of the Optical Society of America A, 4,* 1455–1471.

Hoffman, D. D. (1982). Inferring local surface orientation from motion fields. *Journal of the Optical Society of America, 72,* 888–892.

Hoffman, D. D. & Bennett, B. M. (1985). Inferring the relative three-dimensional positions of two moving points. *Journal of the Optical Society of America A, 2,* 350–353.

Horn, B. K. P. & Schunk, B. G. (1981) Determining optical flow. *Artificial Intelligence, 17,* 185–203.

Koenderink, J. J. & van Doorn, A. J. (1986) Depth and shape from differential perspective in the presence of bending deformations. *Journal of the Optical Society of America A, 3,* 242–249.

Krauskopf, J. (1980) Discrimination and detection of changes in luminance. *Vision Research, 20,* 671–677

Landy, M. S., Dosher, B. A. Sperling, G. & Perkins, M. E. (1988) The kinetic depth effect and optic flow: II. Fourier

and non-Fourier motion. *Mathematical Studies in Perception and Cognition, 85-4*. NYU Report Series.

Landy, M. S., Sperling, G., Dosher, B. A. & Perkins, M. E. (1987). From what kind of motions can structure be inferred? *Investigative Ophthalmology and Visual Science (Supplement), 28*, 233.

Landy, M. S., Sperling, G., Perkins, M. E. & Dosher, B. A. (1987). Perception of complex shape from optic flow. *Journal of the Optical Society of America A: Optics and Image Science, 1987, 4*, No. 13, P95.

Limb, J. O. & Murphy, J. A. (1978). Estimating the velocity of moving images in television signals. *Computer Graphics and Image Processing, 4*, 311–327.

Lucas, B. D. & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *Proceedings of Image Understanding Workshop*, 1221–1230.

Marr, D. & Ullman, S. (1981). Directional selectivity and its use in early visual processing. *Proceedings of the Royal Society of London, B. 211*, 151–180.

Nakayama, K. (1985). Biological image motion processing: A review. *Vision Research, 25*, 625–660.

Patel, A. S. & Jones, R. W. (1968). Increment and decrement visual thresholds. *Journal of the Optical Society of America, 58*, 696–699.

Prazdny, K. (1987). Three-dimensional structure from long-range apparent motion. *Perception, 15*, 619–625.

Rashbass, C. (1970). The visibility of transient changes of luminance. *Journal of Physiology, 210*, 165–186.

Reichardt, W. (1957). Autokorrelationsauswertung als funktionsprinzip des zentralnervensystems. *Zeitschrift Naturforschung, 12b*, 447–457.

Roget, B. J. & Anstis, S. M. (1975). Reversed depth from positive and negative stereograms. *Perception, 4*, 193–201.

Roufs, J. A. J. (1974). Dynamic properties of vision—VI. Stochastic threshold fluctuations and their effect on flash-to-flicker sensitivity ratio. *Vision Research, 14*, 871–888.

van Santen, J. P. H. & Sperling, G. (1984a). A temporal covariance model of motion perception. *Journal of the Optical Society of America A, 1*, 451–473.

van Santen, J. P. H. & Sperling, G. (1984b). Applications of a Reichardt-type model to two-frame motion. *Investigative Ophthalmology and Visual Science (Supplement), 25*, 14.

van Santen, J. P. H. & Sperling, G. (1985). Elaborated Reichardt detectors. *Journal of the Optical Society of America A, 2*, 300–321.

Short, A. D. (1966). Decremental and incremental thresholds. *Journal of Physiology, 185*, 646–654.

Sperling, G. (1976). Movement perception in computer-driven visual displays. *Behavior, Research, Methods and Instrumentation, 8*, 144–151.

Sperling, G., Landy, M. S. Dosher, B. A. & Perkins, M. E. (1989). The kinetic depth effect and identification of shape. *Journal of Experimental Psychology: Human Perception and Performance*, in press.

Ullman, S. (1979). *The Interpretation of Visual Motion*. Cambridge, MA: MIT Press.

Ullman, S. (1985). Maximizing rigidity: The incremental recovery of 3-D structure from rigid and non-rigid motion. *Perception, 13*, 255–274.

Wallach, H. & O'Connell, D. N. (1953). The kinetic depth effect. *Journal of Experimental Psychology, 45*, 205–217.

Watson, A. B. (1986). Temporal sensitivity. In *Handbook of Perception and Human Performance, Volume I: Sensory Processes and Perception* (K. R. Boff, L. Kaufman & J. P. Thomas, Eds). New York: Wiley.

Watson, A. B. & Ahumada, A. J. Jr (1983). A look at motion in the frequency domain. *NASA Technical Memorandum 84352*.

Watson, A. B. & Ahumada, A. J., Jr (1984). A model of how humans sense image motion. *Investigative Ophthalmology and Visual Science (Supplement), 25*, 14.

Watson, A. B. & Ahumada, A. J., Jr (1985). Model of human visual-motion sensing. *Journal of the Optical Society of America A, 1*, 322–342.

Watson, A. B., Ahumada, A. J., Jr and Farrell, J. E. (1986). Window of visibility: A psychophysical theory of fidelity in time-sampled visual motion displays. *Journal of the Optical Society of America A, 3*, 300–307.

Watson, A. B., Thompson, P. G., Murphy, B. J. & Nachmias, J. (1980). Summation and discrimination of gratings moving in opposite directions. *Vision Research, 20*, 341–347.

Williams, D. & Phillips, G. (1986). Structure from motion in a stochastic display. *Journal of the Optical Society of America A, 3*, 30–31.

Williams, D. & Phillips, G. (1987). Rigid 3-D percept from stochastic 1-D motion. *Journal of the Optical Society of America A, 4*, 45.

Journal of Experimental Psychology:
Human Perception and Performance
1989, Vol. 15, No. 4, 826-840

Kinetic Depth Effect and Identification of Shape

George Sperling and Michael S. Landy
New York University

Barbara A. Dosher
Columbia University

Mark E. Perkins
New York University

We introduce an objective shape-identification task for measuring the kinetic depth effect (KDE). A rigidly rotating surface consisting of hills and valleys on an otherwise flat ground was defined by 300 randomly positioned dots. On each trial, 1 of 53 shapes was presented; the observer's task was to identify the shape and its overall direction of rotation. Identification accuracy was an objective measure, with a low guessing base rate, of the observer's perceptual ability to extract 3D structure from 2D motion via KDE. (1) Objective accuracy data were consistent with previously obtained subjective rating judgments of depth and coherence. (2) Along with motion cues, rotating real 3D dot-defined shapes inevitably produced a cue of changing dot density. By shortening dot lifetimes to control dot density, we showed that changing density was neither necessary nor sufficient to account for accuracy; motion alone sufficed. (3) Our shape task was solvable with motion cues from the 6 most relevant locations. We extracted the dots from these locations and used them in a simplified 2D direction-labeling motion task with 6 perceptually flat flow fields. Subjects' performance in the 2D and 3D tasks was equivalent, indicating that the information processing capacity of KDE is not unique. (4) Our proposed structure-from-motion algorithm for the shape task first finds relative minima and maxima of local velocity and then assigns 3D depths proportional to velocity.

In 1953, Wallach and O'Connell described a depth percept derived from motion cues that they called the *kinetic depth effect* (KDE). Since that time, there has been a great deal of research on the KDE, examining the effects of stimulus parameters such as dot numerosity in multidot displays (Braunstein, 1962; Green, 1961), frame timing (Petersik, 1980), occlusion (Andersen & Braunstein, 1983; Proffitt, Bertenthal, & Roberts, 1984), the detection of nonrigidity in the three-dimensional form most consistent with the stimulus (Todd, 1982), and veridicality of the percept (Todd, 1984, 1985).

Since 1979, there have been numerous attempts at modeling how observers and machines could derive three-dimensional (3D) structure from two-dimensional (2D) motion cues. Ullman (1979) referred to this computational task as the *structure-from-motion* problem. Ironically, Ullman's model and most ensuing ones do not explicitly use motion cues. These models are essentially geometry theorems concerning the minimal number of points and views needed to specify the shape under various simplifying constraints such as assumed object rigidity and assumed parallel perspective (Bennett & Hoffman, 1985; Hoffman & Bennett, 1985; Hoffman & Flinchbaugh, 1982; Ullman, 1979; Webb & Aggarwal, 1981). From the geometric models, iterative models have been developed that use newly arrived position data, not to derive the true structure, but to improve the current 3D representation in the sense of maximizing its rigidity (Landy, 1987; Ullman, 1984). Only a few models actually use point velocity (i.e., an optic flow field) in addition to point position (e.g., Clocksin, 1980; Koenderink & van Doorn, 1986; Longuet-Higgins & Prazdny, 1980), and one model also uses point acceleration (Hoffman, 1982).

It has been difficult to relate models of the KDE to the results of psychological studies. An important component of the problem has been the difficulty of finding an appropriate experimental paradigm. Many KDE experiments have used subjective ratings of "depth" or "rigidity" or "coherence" as the responses (see Dosher, Landy, & Sperling, 1989, for a review). Relating subjective responses to a process model of KDE is problematic. Typically, a structure-from-motion model yields a shape specification. To link the derived shape to subjective judgments, and thereby to experimental results, a decision-making apparatus to predict judgments is needed, and this may be quite complex.

Objective Measurements of KDE: Problems

Because the ability to derive structure from motion presumably evolved to solve an objective environmental problem a better approach to studying KDE is to measure the accuracy of the KDE in an objective fashion. Does the observer perceive the correct shape in a display? The correct depths? The correct depth order? The correct curvature? Some of the studies cited earlier attempted to answer such questions by using objective response criteria (e.g., percentage correct in a one- or two-interval forced-choice task). Unfortunately, in almost every case, subjects can achieve good performance on the task by

The work described in this article was supported by The Office of Naval Research, Grant N00014-85-K-0077, and by the U S Air Force Life Sciences Directorate, Visual Information Processing Program Grants 85-0364 and 88-0140.

Correspondence concerning this article should be addressed to George Sperling, Psychology Department, New York University, 6 Washington Place, Room 980, New York, New York 10003.

Atch-17

neglecting perceived depth and consciously or unconsciously formulating their responses on the basis of other cues. In these cases, there is a simple non-KDE cue sufficient to make the judgment accurately. Although the subject may not consciously be using these artifactual cues to make correct judgments, we cannot be sure of the basis of the response until the artifactual cues have been eliminated or rendered useless (e.g., through irrelevant variation).

Let us consider some examples. Lappin, Doner, and Kottas (1980) presented subjects with a two-frame representation of dots randomly positioned on the surface of an opaque rotating sphere displayed by polar projection. On the second frame, a small percentage of the dots were deleted and replaced with new random dots. Subjects were required to determine which of two such two-frame displays had a higher signal-to-noise ratio (in terms of dot correspondences). Lappin et al. (1980) interpreted their results in terms of the "minimal conditions for the visual detection of structure and motion in three dimensions" (p. 717), which is the title of their article. Indeed, the signal dots represent two frames of a rigid rotating sphere. But, subjects do not need to correctly perceive a 3D sphere in order to make a correct response. There was no analysis offered of how far a 3D perception could diverge from spherical and still yield the observed accuracy of response. Alternatively, subjects might base their responses on perceived 2D flow fields, judging the percentage of dots in the first frame that have corresponding dots in the second frame. This 2D judgment need not use the entire motion flow field. For example, the 5.6° 3D motion of the sphere corresponds to a small, essentially linear translation in the center of the field. Discriminating signal-to-noise ratios in translations is related to Braddick's (1974) "dmax" procedures for discriminating perceived linear motion; it does not necessarily have anything to do with KDE. Thus, although Lappin et al. used response accuracy as their dependent variable, the subject's ability to estimate a signal-to-noise ratio may have been artifactual and certainly is not easily converted into an estimate of the accuracy of KDE.

Petersik (1979, 1980) represented rotating spheres by surface elements that were dots or small vectors. In both studies, the spheres were displayed with polar projection, and subjects were required to discriminate clockwise from counterclockwise rotation. A possible artifact here is that the motion of a single stimulus element provides sufficient information to respond correctly. That is, under polar perspective, stimulus points follow elliptical paths in the image plane. To determine rotation direction, the subject needs only determine the 2D rotation direction of a single point (assuming knowledge of the vertical position of the point with respect to eye level). Petersik made the task more difficult by adding noise to some dot paths, by varying the slant of vector elements from frame to frame, or by varying the numerosity. However, none of these manipulations prevents the subject from using a purely 2D, non-KDE strategy. Indeed, Braunstein (1977) had previously examined precisely this point. Braunstein demonstrated that only the vertical component of the polar perspective transformation was used by subjects for a depth-order judgment, and that this component was sufficient.

Andersen and Braunstein (1983) also used discrimination of rotation direction to evaluate KDE. Their displays represented clumps of dots on the surface of a sphere. A clump was construed as being bounded by an invisible pentagon, whose presence was made known by the fact that, when it lay on the front surface of the sphere, it occluded dots that lay behind it on the rear surface. These spheres were displayed by parallel perspective, and the cue to depth order (front, rear) was provided by occlusion. Again, although the dependent variable was response accuracy, a subject did not need to perceive a 3D object to determine the direction of rotation—the subject needed only to determine the movement direction of the continuously visible clumps.

In several studies, simple relative velocity cues are all that the subject needs to perform the KDE task. Braunstein and Andersen (1981) displayed a multidot representation of a dihedral edge that moved horizontally. The dots were displayed using polar projection, so that horizontal point velocities were inversely proportional to depth. Thus, the display contained a velocity gradient that either increased or decreased from the midline of the display to the upper and lower edges of the display. Subjects judged whether a given display represented a convex or concave edge. In this task, comparing the relative velocity of points in the center and at the top edge of the display is all that is necessary to perform accurately (the location with the greater velocity is judged "forward").

In experiments by Todd, subjects determined which of five curvatures (Todd, 1984) or slants (Todd, 1985) were depicted in a multidot display. Again, Todd described the task in terms of the perceived 3D object, but accurate performance is possible by comparing the relative velocities of points in just two areas of the display.

In all the studies just cited, the subject could perform the required KDE task by using a minimal artifactual cue. One possible solution to the problem of subjects learning to use artifactual cues is to withhold feedback. The assumption is that, without feedback, the subject will use only perceived 3D shape. This approach has been used extensively by Todd (1982, 1984, 1985). Unfortunately, withholding feedback does not mean that the subject cannot use an alternative perceptual or decision strategy to supplement judgments of perceived KDE depth. One strategy that subjects often adopt without feedback is to adjust their responses so as to respond equally (or nearly equally) often with each of the possible responses. For example, Todd's (1984) procedure is vulnerable to this artifact of strategy. He used surface dots to represent cylinders with five different curvatures. On a given trial, subjects judged which of the five curvatures was presented. As an alternative to perceiving KDE depth, a subject could judge the apparent velocity of dots in the center of the display and use the knowledge of the velocities displayed on previous trials to choose a curvature category. Indeed, subjects are extremely good at estimating the mean velocity and variations from it in a sequence of displays (McKee, Silverman, & Nakayama, 1986). Although the subjects' use of a trivial strategy that estimates just a single velocity per trial may not explain the entirety of Todd's results, it predicts the nearly veridical

character of subject responses and thereby could account for most of the data.

Objective Measurement of KDE: Proposed Solution

The KDE is a perceptual phenomenon that allows subjects to perceive the relative depth of different positions in visual space and hence to infer the shapes of objects in the environment. In all of the experiments we have discussed, the shapes presented were very simple (spheres, cylinders, and planes), and hence simple response strategies would have been effective. None of the experiments discussed above requires the subject to use a perceived 3D shape in order to perform accurately. In all of the studies we reviewed, subjects had the opportunity to use artifactual cues. None of these experiments presented shapes with complexity approaching that seen in the real world in which the ability to compute structure from motion evolved.

In this article, we describe a new method for investigating KDE. Our aim is to provide, instead of the demonstration of KDE by means of perceptual reports (what subjects say they see), a test of perceptual abilities (what complex shape properties subjects can extract from visual flow fields). The task is shape identification, in which on each trial, one of a large lexicon of shapes is presented. Each shape consists of a flat ground with zero, one, or two bumps or depressions. The bumps and depressions vary in position, 2D extent, and orientation. Because of the way the lexicon of shapes is constructed, good performance in the shape identification task requires simultaneous local computation of velocity in many positions of the display and global coordination of the local information.

Experiment 1: Dot Numerosity and Bump Heights

To demonstrate the shape identification method and to investigate its limits, we replicated and extended one of the classic findings in multidot KDE, the dependence of quality ratings (usually combined coherence and rigidity, or "goodness") on dot numerosity (Braunstein, 1962, Dosher et al., 1989; Green, 1961; Landy, Dosher, & Sperling, 1985). Quality of KDE generally has been found to increase with dot numerosity. We investigated the effects of dot numerosity and depth extent on the effectiveness with which subjects used the KDE to identify the target shape from among its many close competitors.

Method

Subjects. Three subjects were used in the study. Two were authors of this article, and the third was a graduate student naive to the purposes of the experiment. Two subjects had normal or corrected-to-normal vision; one subject (CFS) had vision correctable only to 20:40.

Displays. The shapes used in the experiment were 3D surfaces consisting of zero, one, or two bumps or concavities on an otherwise flat ground. Here we use the term *shape* to indicate the positions of these bumps and concavities on the flat ground, irrespective of other stimulus parameters that were varied, including bump height, number of dots used to represent the shape, and rotation direction. The shapes were constructed as follows (see Figure 1A). Within a square area with sides of length s, a circle with diameter $0.9s$ was centered. All depth values outside the circle were set to zero (i.e., in the object base plane, which in the initial display was the same as the image plane). For each of three positions inside the circle (located at the vertices of an equilateral triangle), the depth was specified as either $+h$ (a distance h in front of the object base plane, closer to the observer), 0 (in the object base plane), or $-h$ (behind the object base plane). A smooth spline was constructed, using a standard cubic spline algorithm, which passed through the flat surround and the vertices of the triangle. For a given set of vertices, 27 shapes were constructed in this way (see Figure 1B for some examples).

Two different sets of vertices were used to generate shapes. These were either at the corners of a triangle pointing up (designated *u*) or of a triangle pointing down (designated *d*). Shapes were denoted by indicating the trio of positions (*u* or *d*), and then specifying for each position (in the order shown in Figure 1A) whether that position was in front of the object base plane (+), in the plane (0), or behind it (−). For example, the shape denoted by *u*+−0 consists of a bump in the upper central area of the display, a depression in the lower left of the shape, and a flat area in the lower right of the shape (see Figure 1B). Note that *u*000 and *d*000 both designate the same shape: a flat square. Fifty-three distinct shapes can be generated in this manner.

Displays were generated for all combinations of the 53 shapes, three dot numerosities, and three bump heights. For the flat shape (denoted *u*000 or *d*000), varying bump height has no effect, and so there are only three flat shape display types (corresponding to the three numerosities). For all other shapes there are nine display types. This results in 471 display types. For most display types, a single instantiation was generated (choosing a set of random dots and forming a display after rotation and projection). For each of the display types for the flat shape, six instantiations were made. Thus, there were 486 different displays. Bump height, h, was 0.5s, 0.15s, or 0.05s, where s is the length of a side of the square ground. The 3D perspective drawings of the shapes in Figure 1B are for the largest bump heights. Dot numerosities were 20, 80, and 320. The bump height and dot numerosity manipulations are illustrated in Figures 1C and 1D, respectively.

Multidot displays of these shapes were generated by choosing a random sample of positions on each surface, rotating the resulting set of points about a fixed vertical axis, and projecting them onto an image plane via parallel projection. The 3D motion was a single cycle of a sinusoidal rotation about a fixed vertical axis through the center of the object base plane, with amplitude of 25° and period of 30 frames. More specifically, the angle at which the base plane was oriented with respect to the image plane was $\theta(m) = \pm 25 \sin(2\pi m/30)$ degrees, where m is the frame number within the 30 frame display.

Two rotation directions were used, indicated as *l* and *r*, corresponding to whether the left or right edge of the display came forward initially. Equivalently, this described the side of the observer to which the shape "faced" in the second half of the rotation (which was usually an easier way to code the response). For an *l* rotation (see Figure 1E), the object initially appeared face-forward. It was then rotated so that the front moved to the right until the object had rotated 25°. Then it reversed direction and rotated to the left until it was 25° to the left of its initial orientation. Finally, it again reversed direction and rotated until the ground plane was again perpendicular to the line of sight. A full description of a display by a subject included the indication of the set of vertices (*u* or *d*), the 3D depths at these vertices (+,−,0) and the direction of rotation (*l* or *r*), for example, *u*+−0*l*.

Because of the parallel projection, simultaneous reversal of depth signs and of rotation direction yields precisely the same physical image sequence. The 486 displays described earlier were all generated

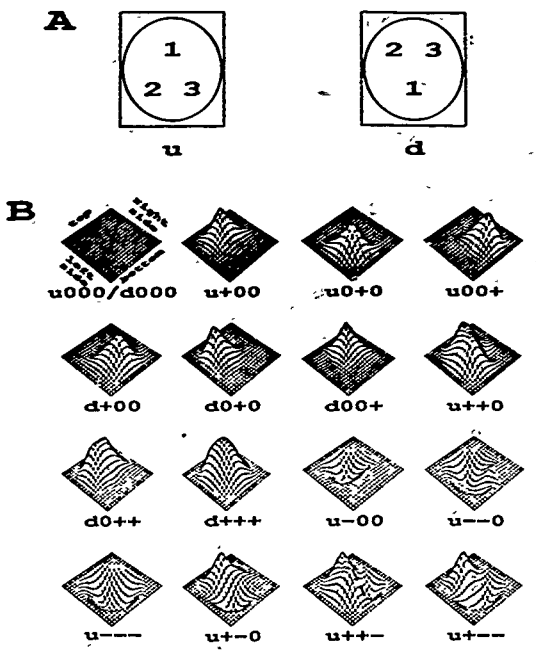


Figure 1. Stimulus shapes, rotations, and their designations. (Shapes were constructed by smoothly splining a flat ground and three points that were either toward the observer [plus sign], in the flat ground [zero], or away from the observer [minus sign].) A: These three points were at the corners of one of two possible equilateral triangles, for which the odd point is up [*u*] or the odd point is down [*d*]. In the experiment, subjects were required to name the shape and rotation direction perceived. The numbers specify the order in which the depth signs of the three points are to be reported. B: The various combinations result in a lexicon of 53 shapes; typical examples are illustrated here as perspective plots. The orientation of these plots relative to the viewing direction is indicated on the first example.

(Figure continues)

with the *l* rotation, but each can equally well be described as an *r* rotation of the sign-reversed shape. There are 108 ways to designate a display by combining an up or down shape-type with a bump, depression, or flat surface at three different locations with a left or right initial direction of motion; that is, $[d, u] \times [+, -, 0]^3 \times [l, r]$. For most shapes, there are two equally valid ways to describe the display. For example, *u*+−0*l* and *u*−+0*r* describe the same display. The flat shape is denoted equally accurately as *u*000*l*, *u*000*r*, *d*000*l*, and *d*000*r*. Given the four instantiations of the flat shape, chance performance depends on subject strategy. Repeated responses of *u*000*l* (and its equivalents) yields a guaranteed performance of 18 in 486 correct (or 2 in 54). Random guessing yields an expected performance of just over 1 in 54 correct. Subjects did not designate bump height in their responses. Except in the case of the flat stimuli, bump height was obvious.

After sampling, rotation, and projection, any given frame of the display consisted of *n* points in the image plane. These points were displayed as bright dots on a dark background. The square image extent of the displays projected to a 182 × 182 pixel area subtending 4° of visual angle. The displays were not windowed in any way, so the edges of the display oscillated in and out with the rotation. With the 25° wiggle, at the instants when rotation reverses, the display has shrunk to 90% of its initial horizontal extent.

Displays were presented on a background that was uniformly dark (approximately 0.001 cd/m^2). Dots were single pixels of approximately 65 μcd and were viewed from a distance of 1.6 m. A trial sequence consisted of a cue/fixation spot presented for 1 s, a 1-s blank interval, and the 2-s stimulus sequence. The stimulus sequence was followed by a blank screen, the luminance of which was the same as the background of the stimulus. The display was run at 60 Hz noninterlaced. Each display frame was repeated four times, for an effective rate of 15 new frames per second. The duration of each 30-frame display was 2 s.

Apparatus. Stimuli were computed in advance of the session and stored on disk. The stimuli were processed for display by an Adage RDS-3000 image display system and were displayed on a Conrac

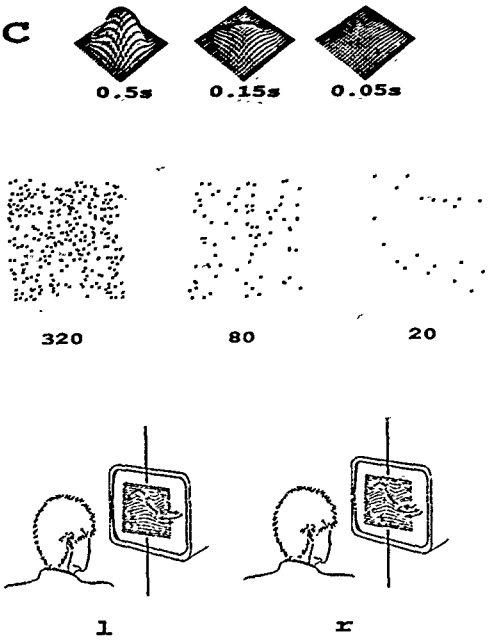


Figure 1 (continued). C: Three bump heights were used: 0.5*s*, 0.15*s*, and 0.05*s*, where *s* is the length of a side of the square base of the shape. The shape depicted here is *u*+++. D. Three dot numerosities were used: 20, 80, and 320. Pictured are the first frames of a representative display in each numerosity condition. E: Two rigid rotation motions were simulated. Both were sinusoidal rotations about a vertical axis through the center of the object ground. The object either first rotated to face the subject's right, then to the subject's left, then returned face-forward [*l*], or in the opposite direction [*r*].

7211C19 RGB color monitor. The stimuli appeared as white dots on a black background.

Viewing conditions. Stimuli were viewed monocularly (with the dominant eye) through a black-cloth viewing tunnel. In order to minimize absolute distance cues, a circular aperture slightly larger than the square display area restricted the field of view. Stimuli were viewed from a distance of 1.6 m. After each stimulus presentation, the subject typed a response on a computer terminal. Room illumination was dim. (Illuminance was approximately 8 cd/m^2.)

Procedure. Subjects were shown perspective drawings of the shapes (as in Figure 1B) and were instructed as to how they were constructed and named. They were told that they would be shown multidot versions of these shapes and would be required to name the shape displayed and its rotation direction as accurately as possible. They were told to use any method they chose to remember and apply the shape and rotation designations.

Each of the 486 displays was viewed once by each subject. The displays were presented in a mixed-list design in four sessions of 45 min each. After each response, the possible correct responses were listed as feedback. For each stimulus, there were always two responses that were scored as correct (given perceptual reversals). For the flat stimuli, four possible answers were correct.

To become familiar with the task and the method of response, each subject ran trials consisting of 27 of the easiest stimuli (the 320 dot 0.5*s*-height stimuli). Subjects ran [illegible] until accuracy was at least 85% correct (approximately 100–1[illegible]

Results

Accuracy data. All subjects reported that they perceived a 3D surface the first and every subsequent time they viewed the high numerosity displays. With low numerosities, the dots were perceived in approximately their correct positions in 3D space, but there were too few dots to give the illusion of a continuous surface or to discriminate unambiguously between alternative responses. The very limited practice served merely

to teach the subjects to name the perceived shapes without having to refer to drawings.

The results of Experiment 1 are summarized in Figure 2. Each response was scored as correct only if both the shape and the rotation direction were correct and consistent. Thus, if *u*+−0/ was the display, responses *u*+−0/ and *u*−+0r were correct. Every other response was incorrect. There were occasional responses with the correct shape and the incorrect rotation direction (66 such errors, 4.5% of all responses, 10% of all errors). Subjects later indicated that most of these were a result of forgetting the direction of rotation before the response was completed, rather than from a truly misrotating percept. Nevertheless, such responses were treated as incorrect.

As expected, accuracy improved both with the numerosity and with the amount of depth displayed. There were signs of a ceiling in performance as numerosity increased. For two subjects, for 320 point displays, the curves crossed, and the middle-range depth extent (0.15*s*) was as good or better than the large 0.5*s*-depth extent. An analysis of variance was computed treating numerosity, height, and subjects as treatments, and shapes/rotations as the experimental units. Both numerosity and degree of depth were highly significant ($p < .0001$), with $F(2, 106) = 119.0$ and $F(2, 106) = 102.9$, respectively. Subjects differed significantly from one another, $F(2, 106) = 33.5$, $p < .0001$. The three-way interaction was significant, $F(8, 424) = 2.6$, $p < .01$, indicating that the interaction of height and number differed among subjects (see Figure 2). No two-way interactions were significant.

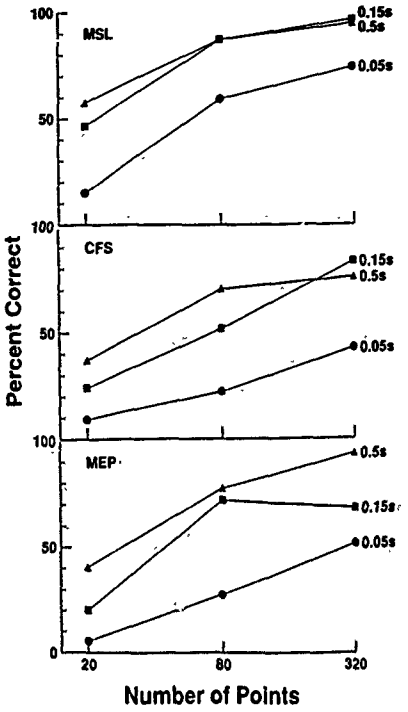


Figure 2. Performance on the shape identification task as number of points in the simulated shape was varied. (The parameter is the height of the bumps relative to the length of a side. Each panel represents data from a different subject. Performance increased with both numerosity and bump height.)

Error analyses. A confusion matrix was computed, pooled across subjects, the nine conditions, two rotation directions, and two possible designations of each shape or depth reversals (it was thus a 27 × 27 = 729 cell matrix). Table 1 is a summary of these identification errors. Descriptions are given for seven common error types, one uncommon error type and a miscellaneous category. If a bump and a depression were present in the display, and only one of the two was indicated by the subject, this was called a *missed feature error*. If the bump and depression are of equal extent on the base plane (e.g., *u*+−0), then this was called a *missed equal size feature*. If they were of unequal extent, and the smaller of the two was not reported, this was categorized as a *missed smaller feature*. Any display that contained only one depth sign (such as *u*+00) and was reported as containing both depth signs (e.g., *u*0+−) was categorized as *report two depth signs when there was only one*. For any given row in the table, the second column presents examples of errors of that row type. The third column lists the number of cells in the confusion matrix that correspond to an error of a given type, and the fourth column provides the total number of errors that occurred over all cells of that type. The last column is the average number of errors per cell in cells of that type, computed as the ratio of the number of trials indicated in Column 4 divided by the number of cells in Column 3. In total, there were 586 errors; divided by 702 error cells this yields 0.83 errors per cell on the average. A ratio greater than 0 83 in Column 5 of Table 1 indicates an error type more common than the average, a smaller number indicates a less common than average error type.

The bottom row of the table provides summary information. The first seven error types listed had ratios well over this value and thus were more common than other errors. The *report two depth signs* ... error type is an example of an exceedingly uncommon error.

The quantity of data collected was not sufficient to enable us to confidently draw many specific conclusions from the error data. The hypothesis that errors are distributed uniformly across the nine error classes was easily rejected, $\chi^2(8, N = 586) = 1{,}032$, $p < .001$. It appears that four types of errors were the most prevalent. Large single bumps were highly confusable, especially the subtle difference in shape that distinguishes *d*+++ from *u*+++, but also that distinguishes between *d*+++ and *d*0++, and so on. Errors were made in horizontal location of the shape within the ground (e.g., *d*0+0 was reported as being *u*+00, or *d*++0 as *u*+0+). Errors were also made in judging the width of the bumps

Table 1
Summary of Identification Errors, Pooled Over Subjects, Bump Heights, Dot Densities, Rotation Directions, and Depth Reversals

Description	Examples	Number of cells	Number of errors	Ratio*
Small distortions of large bumps	*u*+++ interchanged with *d*+++	2	29	14.5
Incorrect bump width, correct location	*u*0++ interchanged with *d*+00	4	34	8.5
Missed smaller features	*u*++– reported as *u*++0	6	30	5 0
Diagonal bump reported as large bump	*u*++0 reported as *u*+++ or *d*+++	8	23	2.9
Missed equal size feature	*u*+0– reported as *u*+00	12	29	2.4
Incorrect diagonal bump size	*u*++– reported as *u*+0–	8	16	2 0
Small horizontal location error	*u*+00 interchanged with *d*0+0	16	27	1.7
Report two depth signs when there was only one	*u*+00 reported as *u*+–0	168	40	0.24
Other errors	—	478	358	0.75
All errors	—	702	586	0.83

*Total number of indicated error responses divided by total number of applicable cells (Column 4/Column 3). A ratio greater than 0 83 indicates a type of error that is more common than average.

(e.g., *d*+00 was reported as *u*0++). Finally, for displays for which both a bump and a concavity were present, occasionally one of the two was not noticed. It is interesting to note that in every case of this type of error (the missed smaller features and missed equal-size features of Table 1, and the less common missed larger features), the response was of a single bump toward the observer. In other words, in the presence of a perceived convexity, a concavity is occasionally missed, but not the other way around. On the other hand, when only one nonzero depth was present (a single bump or concavity), it was very rare for subjects to give a response containing multiple depth signs.

When the confusion matrix was broken down by experimental condition, the amount of data was rather low. Nevertheless, a few interesting trends were evident. First, all seven common error types (the first seven rows of Table 1), remained common in all experimental conditions. As the task became more difficult, the types of errors subjects made remained "sensible." Second, the first two error types, although common in difficult conditions (low height or low numerosity), became even more common in easier conditions. As the shape impression improved, the subjects were able to eliminate other possible shapes and then were more likely to err by choosing the most similar incorrect shape. The distinction between *d*+++ and *u*+++ was very difficult to make even when the perception of depth was quite compelling and well sampled. The *report two depth signs* . . . error type was uncommon in all conditions, but there appeared to be a trend for this error type to become more common as numerosity increased.

Experiment 2: Texture Density

Several cues may lead to correct shape identification in the KDE task. One cue is dynamic changes in texture density. The shapes are generated in such a manner that, head-on (i.e., viewed with the object base plane in the picture plane), the expected local dot density across the display is uniform. By itself, the initial frame has no shape information whatsoever. As the shape rotates, areas in the display become more dense or sparse as the areas in the shape that they portray become more or less slanted from the observer. Theoretically, the observer could use this cue from subsequent frames after the first to determine the shape. Because we are interested in structure from motion, the changing texture density adds a cue in addition to the relative motion cue. In Experiment 2, we compared three conditions: (a) Both the motion and density cues were present as before; (b) only the motion cue was present—dot lifetimes were varied in such a way as to eliminate the density cue by keeping local average dot density constant across the display; and (c) only the density cue was present—the relative motion cue was eliminated by reducing dot lifetimes to just one frame.

Method

Subjects. Three subjects were used in the study. One was an author of this article; two were graduate students naive to the purposes of the experiment. Two had corrected-to-normal vision; one subject (CFS) had vision correctable only to 20.40.

Displays. The displays were generated in a manner similar to Experiment 1. The same lexicon of 53 shapes was used. The flat ground surrounding each shape was extended horizontally by 20% and was later windowed to the same 182 × 182 pixel, 4° square, so that the sides of the displays no longer oscillated with the rotation. Instead, points appeared and disappeared at the edges of the window. For each shape, an instantiation of the shape was made with 10,000 points and with the large 0.5s-bump height of Experiment 1. Displays for each of the three experimental conditions were made by randomly subsampling points from this rotating 10,000-dot shape.

Control condition: Motion and texture cues. The control condition had both the relative motion and changing texture density cues. A small random subsample of points was chosen, so that approximately 320 points were visible through the 4° square window. The subsample of points was rotated and projected as before, and then clipped so that only those points within the window were displayed. This condition was identical to the easiest condition of Experiment 1 (0.5*s*, 320 dots) except for the windowing (and the lower dot contrast described later). Examples of the density cue available in these displays are shown in Figure 3.

Only motion cue. This main experimental condition removed the changing texture density cue (Figure 3). The 4° × 4° square window was treated as consisting of a 100 × 100 grid of subsquares. Texture density was kept uniform by forcing each subsquare to contain exactly 3 points in every display frame. Thus, there were exactly 300 points visible in every frame. On the first frame, 300 of the 10,000 points were randomly chosen, subject to the constraint that exactly 3 points were chosen for each subsquare. On each subsequent frame, the 10,000 points were sorted by the proper amount. Then, for each of the 100 subsquares, the points (of the 300) that then appeared in each subsquare were counted. If more than three occurred, points were randomly chosen and marked as no longer displayed, until the number of displayed points in that subsquare fell to 3. If less than 3 points in a grid square were displayed, then more points were randomly chosen (from the 10,000) that would then appear in that subsquare to bring the total back up to 3. In this condition, dot density remained uniform throughout the display. Points were deleted or reinstated only as needed to keep the density uniform. Although variations in texture density were noticeable in the control displays, the exclusion of the density cue did not seriously disrupt the correspondence of the majority of the points: Most points remained displayed for 10 frames or more during the 30 frame display.

The amount of scintillation was small. The average change (one half of total dot additions plus deletions) between two frames was 16; for 300 dot displays this was 5.3% scintillation. (The highest between-frame scintillation was 8 3%:)

Only texture density cue. The relative motion cue was removed in this condition leaving the changing texture-density cue intact. For each frame in the display, 320 of the 10,000 points were randomly chosen. This happened independently on every single frame, subject to the constraint that no point ever appeared in two successive frames. Thus, no relative motion cues were available in these displays, which looked like dynamic sparse random dot noise. On the other hand, because the points were chosen randomly from the 10,000 points, they had the same expected texture density as the 10,000 points on each frame, and indeed became more dense and sparse in exactly the same fashion as in the first experimental condition (as illustrated in Figure 3).

There were 53 possible shapes and three experimental conditions, resulting in 159 display types. Two different displays were made of each display type of the flat shape, and one display was made for all other display types. There were thus 162 displays. They were displayed as bright green dots on a green background of lesser luminance. The display background luminance was 31 cd/m^2. Each dot added an additional 13 μcd, viewed from a distance of 1.6 m. All other display characteristics were the same as in Experiment 1.

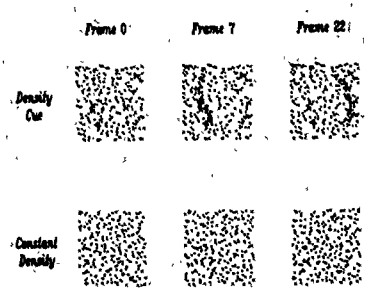


Figure 3. The dynamic density cue. (Three frames are shown from a display corresponding to $u+0+r$, a bump extending from the top center to the lower right. The upper row shows frames with the density cue. The lower row illustrates the effectiveness of removing the density cue in the motion-only condition.)

Apparatus. The apparatus was the same as in Experiment 1. Only the green channel of the Conrac display monitor was used.

Viewing conditions. The viewing conditions were identical to Experiment 1.

Procedure. There were 11 experimental conditions, the 3 described previously (motion and texture, motion only, texture only) and 8 others that will be reported elsewhere. There were thus a total of 594 displays, including the 162 displays of the 3 conditions reported here. These were presented in a mixed-list design in four sessions of 1 hr each. Otherwise, the procedure was identical to Experiment 1.

Results

Density cue. The results are shown in Figure 4. For two subjects (MSL and CFS), elimination of the changing density cue did not alter performance. For the third subject (JBL), performance dropped from 81.5% to 68.5% after the density cue was eliminated. However, it was not clear whether this small performance change was due to the elimination of the density cue itself or the introduction of scintillation (dot noncorrespondences) by the process of eliminating density cues. For two subjects (CFS and JBL), the elimination of the relative motion cue in the density only condition dropped performance to levels that did not differ significantly from chance. For the third subject (MSL), performance with the density cue alone was significantly above chance, although well below performance for conditions in which the relative motion cue was available.

In the condition in which only the changing dot density cue was available, the displays did not look 3D. The only subject (one of the authors) who was able to perform significantly above chance in this condition was highly familiar with the construction of the displays. For any given shape and rotation direction, clumps of higher density appeared first on one side of the display, and then later on the other side, as the object was rotated an equal amount in both directions from the initial face-forward orientation through the course of the 30-frame display. Performance was a matter of noting the positions in the display at which high density occurred,

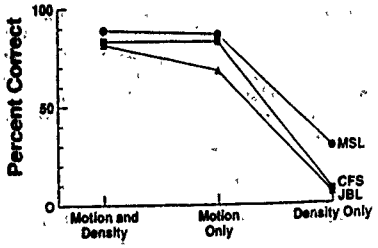


Figure 4. Percentage of correct shape-and-rotation identifications for the three cue conditions of Experiment 2. (Data are shown for 3 subjects.)

on which side of the display they occurred first, and the 2D shape of the texture clump. Then, a response was chosen that was most consistent with this information. This was a highly cognitive task, and it took far longer to respond in this condition as a result.

Changing dot density was neither a necessary nor a sufficient cue for the perception of 3D shape with these displays. However, when the density cue was available with motion cues, the density cue may have been used by one of three subjects to slightly improve his responses. When the density cue was the only cue, another one of three subjects was able to improve his response accuracy to significantly above chance. These results point out the importance of removing artifactual cues from kinetic depth displays.

Scintillation cue. In the constant density condition, one might argue that the subject was indirectly provided with shape information by the amount of scintillation (dot noncorrespondence) in different areas of the display. Local scintillation could potentially be used by a subject (just as density information was useful to one of three subjects in the density-only condition).

The relation between local scintillation in these displays and local density (and thereby, ultimately, local shape) in the control displays is not simple. Points are deleted or added only when necessary to keep the number of points in a given locale constant. The number of points that will be added (or deleted) is thus proportional to the local rate of change of texture density. The difficulty in computing shape from scintillation is that subjects are poor at judging the degree of scintillation in a pattern, other than differentiating some scintillation from no scintillation (Lappin et al., 1980). And it is even more difficult to determine whether scintillation is due to points being added or to points being subtracted, that is, to determine the sign of the change of texture density.

We further investigated the possibility that scintillation might have been a useful cue, in an informal experiment. Various amounts of irrelevant scintillation (in the form of fresh, randomly occurring dots in each frame) was added to all areas of each frame. With added scintillation that was 10 times more than that produced by the density removal program, the quality of the image was greatly impaired. But the ability to discriminate shapes seemed to be unimpaired. This means that scintillation is relatively unimportant: Large amounts do not greatly impair the display; small amounts are not necessary to perceive KDE because, when they are masked by large amounts of scintillation, performance hardly suffers.

In displays similar to those of Experiment 2, restricting dots to have lifetimes of only 3 frames was another operation that generated large amounts of scintillation. KDE identification performance remained high even though the amount of scintillation was large and varied randomly throughout the display and from frame to frame (Dosher, Landy, & Sperling, in press; Landy, Sperling, Dosher, & Perkins, 1987). All in all, the difficulty subjects had in estimating the amount of scintillation in the first place and the subsequent difficulty of any computation for estimating shape from scintillation made it unlikely that scintillation played a significant role. We conclude that density-related shape cues are eliminated in the motion-only displays.

Experiment 3: Equivalent 2D Task

Because of the large set of shapes, the systematic way in which it was constructed, and the large set of possible responses, it appears difficult to perform accurately in this task without a global perception of shape. Indeed, except in the case of the density-only displays of Experiment 2, all of our subjects reported perceiving a global shape and basing their response on this global shape percept. Nevertheless, one of our most serious objections to previous studies of KDE was that the subjects could have performed the experimental tasks without a global perception of shape by using minimal, incidental cues. Because our set of shapes was finite (53 shapes), there were indeed potential artifactual strategies; however, because each realization of a shape was composed of different random dots, we were unable to discover any simple, minimal computation for our task. The simplest computation was equivalent to what we believe the KDE computation itself to be.

To study alternative mental computations that might yield correct responses in our KDE task, we developed a new display that did not produce the 3D depth percept of KDE but that was as equivalent as possible to the KDE display in other respects. To perform correctly with the new display, the subject would have to perform a computation that was equivalent to the KDE computation except in that it is performed by some other perceptual/cognitive process, a process that did not yield perceptual depth. We call such a computation a *KDE-alternative computation.*

Suppose that a subject chose to perform the shape identification task by measuring instantaneous velocities at only a small number of spatial positions and making this velocity determination at only a single moment during the motion sequence, for example, a moment at which velocities were the greatest. A high velocity indicates a point far forward or far behind the base plane. Opposite velocities indicate points at opposite depths. Using these simple principles, it is obvious that velocity measurements at six positions, the corners of both triangles used in specifying the shapes, would be sufficient to identify the shapes. Fewer measurements of velocity made at intermediate points would suffice for identification of our restricted set of stimuli, but they would involve unrealistically complicated computations that were specific to this stimulus set.

In Experiment 3, we evaluated a computation for shape reconstruction based on a strategy of making six simultaneous local velocity measurements at the points that corresponded to the possible depth extrema in our stimulus set.

Method

Choosing motion trajectories for display In the shape identification task (Figure 1), suppose one were to track a single point on the surface of the shape throughout the course of the display. Initially the point is at position (x, y, z), where x and y are the horizontal and vertical image plane axes, respectively, and z is the depth axis. As in Experiments 1 and 2, assume that the shape is rotated about the y axis according to $\theta(m) = \pm 25 \sin(2\pi m/30)$, where m is the frame

number. Under parallel projection, the motion path of the point is purely horizontal:

$$x(m) = r \cos\left\{\frac{2\pi}{360}\left[\theta_0 \pm 25 \sin\left(\frac{2\pi m}{30}\right)\right]\right\},$$

where $r = (x^2 + z^2)^{1/2}$, and $\theta_0 = \tan^{-1}(z/x)$ degrees

If the subjects were to apply the local motion strategy to the shape identification task, they would need to measure and categorize local velocity for six such motion paths simultaneously. In Experiment 3, the subjects were presented directly with stimuli containing six moving patches and they were requested to categorize the local directions of motion.

Displays. Each display was based on a particular shape from the shape identification task. Each of the six motion paths portrayed in the display was based on a motion path followed by a critical point on the surface of the shape, as just described. The six critical points were the projections onto the surface of the six points originally used to generate the shapes (see Figure 1A, *u* and *d*). The motion paths were based on the shapes with the largest heights ($h = 0.5s$, where s is the width of the visible background plane).

The displays were intended to force subjects to use the strategy of simultaneously measuring six velocities, without any possibility of recourse to using perceived 3D shape. Each display consisted of six patches of moving random dots (Figure 5). The dots within a patch all moved with the same velocity, and patches were spatially separated, so that there was no perception of depth. The outline squares of Figure 5 were not directly visible to the subject. They acted as windows through which planes of moving random dots were seen. Due to a setup error, dot density in Experiment 3 was slightly less (0.83 of rather than equal to) than the density used in the constant density condition of Experiment 2. (This density difference was so small that it went unnoticed at the time.)

Response mapping. There were two rows of three patches of moving dots. Figure 5 indicates the correspondence of patch position to where that patch's motion is visible in the original shape displays. Spatial positions in Experiments 2 and 3 were essentially similar

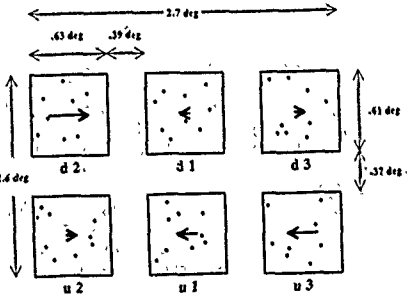


Figure 5 Spatial layout of the stimuli used in Experiment 3. (The squares represent windows through which fields of moving random dots were seen. The outline of the windows was not visible to the subject. The label under each window denotes the position in the shape, as in Figure 1A, that controlled the motion portrayed in that window. For example, the motion path of all the random dots seen in the upper middle window was the same as that taken by the point in a shape display of Experiment 1 that was initially above position *l* in the *d* triangle shown in Figure 1A.)

except that the middle positions in each row of Experiment 2 displays were interchanged to create the Experiment 3 displays. This was done in order to make the response easier for the subjects. With the KD-shape displays, the subject decided whether the three important points were those of the *u* or *d* triangle, and then categorized the height at each of the three corners of that triangle. In the corresponding motion task, the subject decided whether the top or bottom row of patches was most important, and then categorized the motion path of each patch in that row.

For points at a reasonable height above the base plane, the 2D motion path was quasisinusoidal. That is, points moved to the left, then to the right, then returned leftward to their starting position (or right, then left, then right). Points with a larger initial z value moved faster. The extreme z values generated the highest speeds, and these always lay above the vertices of the base triangle used to generate the shape. This meant that subjects could solve the motion task by first judging which row contained the fastest speed, and then, for that row, categorizing the motion in each of the three patches about halfway through the course of the display time. Each patch was to be labeled as moving quickly to the left (*l*), quickly to the right (*r*), or slowly, if at all (0). Note that points in the other row also moved in a quasi-sinusoidal manner, but more slowly than the maximum speed in the relevant row.

One possible response was, for example, *ulr*0. This response would indicate that the fastest speeds were in the upper row; the upper-left patch moved right, then left, then right, the upper-middle patch moved left, then right, then left, and the upper-right patch was moving slowly. There were 54 possible responses (2 rows, 3 possible motion categories for each of the three patches in that row). Because *u*000 and *d*000 denoted the same display (one in which all patches were moving slowly), this yielded 53 distinct display types, corresponding to the 53 distinct shape-and-rotation display types in the shape-identification experiment.

There were 53 possible shapes. With 2 exemplars of the flat shape, and 1 for all other shapes, this yielded 54 displays. Motion displays were displayed as bright white dots on a gray background. The display background luminance was 15.6 cd/m^2. Each dot added an additional 24.3 μcd, viewed from a distance of 1.6 m. All other display characteristics were the same as in Experiment 1.

Apparatus. The apparatus was the same as in Experiment 1, except that a monochrome U.S. Pixel PX15H3151.HS monitor with a fast, white phosphor was used.

Viewing conditions. Stimuli were viewed monocularly with goggles; a circular aperture restricted the field of view. Luminance outside the aperture was approximately equal to the background luminance on the CRT, which was 15.6 cd/m^2. Stimuli were viewed from a distance of 1.6 m. After each stimulus presentation, the subject keyed responses using response buttons, and visual feedback was given on the CRT. The room was dark, but light adaptation level was controlled by the CRT background and the illumination of the occluding screen.

Procedure. A block of trials consisted of 108 trials. Each of the 54 displays was viewed twice in random order. For the stimuli based on the flat shape, two possible answers were correct (*u*000 and *d*000). For all other stimuli only one answer was correct.

Subjects were told precisely the correct strategy to use. They were told that they would see six patches of moving dots. They were to determine which row contained the patches with the fastest motion (either the upper row, designated *u*, or the lower row, designated *d*). For that row, subjects were to categorize the motion in each of the three patches in that row as measured about halfway through the course of the display time. Each patch was to be labeled as moving quickly to the left (*l*), quickly to the right (*r*), or slowly if at all (0). After each response, the correct answers were displayed as feedback. Other details of the procedure were identical to Experiment 1.

Subjects. Two subjects were used in the study. One was an author of this article, one was a graduate student naive to the purposes of the experiment. Both had corrected-to-normal vision. Subject MSL ran a single block of 108 trials. Subject JBL ran three blocks of 108 trials.

Results

Subject MSL scored 90.7% correct on a single block of 108 trials. In the three blocks of trials run by subject JBL, the scores were 58.3%, 75.9%, and 88.0%, respectively. Indeed, after a little practice, performance was quite good, equal or slightly better than performance in the easiest conditions of Experiments 1 and 2, which had a comparable dot density and range of velocities.

There were too few trials to make an in-depth analysis of error data. However, the most frequent motion response errors corresponded to the two most frequent KDE errors in Table 1 (small distortions or mislocalizations of large bumps). For example, 8 out of the 10 errors made by MSL were analogues of these two error types. Examples: *ulll*, a triple "up" bump was reported as *dlll*, a triple "down" bump; *u0l* was reported as *d0l*; a double bump was mistaken for a single bump in the same location (see Figure 1). Indeed, these results are not surprising because the velocities involved in Experiments 1 and 3 were similar. It seems likely that a very large number of trials would be required to find any significant differences in the error patterns in Experiment 3 and those in Experiment 1.

Discussion

We have introduced a new objective task for measuring the perceptual effectiveness of the kinetic depth effect: shape identification. With the current lexicon of shapes, it measures whether the subject can globally determine precisely which areas are in front of the ground and which areas are behind the ground. We consider here some possible objections to and some issues raised by our results.

Cues to Structure From Motion: Optic Flow or Interpoint Distances?

In the displays of Experiment 2, in which dot density was controlled, subjects solved the shape identification task even though no single frame contained any information that could have been used to infer shape. For these stimuli, at least two frames were needed to infer shape. By definition then, the only possible cues were motion cues.

There are at least two possible motion cues to depth: optic flow and changing interpoint distances in the displays. That is, subjects could be deriving shape from a global optic flow field (instantaneous velocity vector measurements across the field) or from measurement of interpoint distances of particular dots over two or more frames. Models of the KDE have been based on both optic flow (Koenderink & van Doorn, 1986) and on interpoint distances (Hildreth & Grzywacz, 1986; Landy, 1987; Ullman, 1984). To a certain extent, it is possible to differentiate between these models by creating displays in which dots have lifetimes of only two frames. In such displays, a global optic flow field is available (although noisy), and 3D structure could, in principle, be computed from the flow field. Alternatively, some subset of the points could have been used to compute a 3D object based on interpoint distances. However, the particular object changes rapidly because within two frames all points have been replaced by entirely new points, uncorrelated with those of the preceding frames. It turns out that subjects are quite adept at the shape identification task with such displays (Dosher et al., in press; Landy et al., 1987). This, and related results, are taken as strong evidence against the interpoint distance models (Dosher et al., in press; Landy et al., 1987). Together with the results of the present experiment, in which changing density is eliminated as an alternative, this leaves motion flow fields as the necessary and sufficient cue for KDE in moving-dot displays. Whether interpoint distances or other motion cues are ever perceptually salient remain open questions.

Multiple Facets of the KDE

We have previously argued (Dosher et al., 1989; Landy et al., 1985) that measurement of the full effect of stimulus manipulations on the KDE requires several subject responses in order to describe fully the richness of the percept. These responses included judgments of coherence (whether the multidot stimulus coheres as a single object), rigidity (does the object stretch?), and depth extent (what is the amount of depth perceived?). These different aspects of the percept are partially correlated, but they can be decoupled by suitable display manipulations. For example, with some subjects, the addition of exaggerated polar perspective to a display increases the perceived depth extent even as it decreases perceived rigidity.

In the current experiments, this richness of the KDE percept was not explored. We measured the extent to which the display was effective in creating a global sensation of depth, and hence supported objective shape identification. Other aspects such as depth extent or rigidity were not measured. The difference between the three depth conditions was immediately obvious to subjects, and increasing the depth extent displayed (within certain limits) did improve performance, but we did not measure perceived depth extent.

Although perceived rigidity was not explicitly measured, nonrigid percepts were spontaneously reported by subjects. One particular example was very common. Shapes with both bumps and concavities (e.g., *u++−*) were occasionally seen in a nonrigid mode. Rather than seeing one area forward, another one back, and the whole thing rigidly rotating, observers perceived both areas as being in front of the object ground and rotating in opposite directions (this percept looks rather like a mitten with the thumb and finger portions alternately grasping and opening). This particular nonrigid percept occurred most often when the number of dots was large and the depth extent was at its largest. In this stimulus condition, with mixed-sign shapes, it is clearly visible that the two bumps cross (in the rigid mode, one sees through the bump to the concavity behind it when they cross). This is an example of a failure of the "rigidity hypothesis" (Adelson,

1985; Braunstein & Andersen, 1984; Dosher, Sperling, & Wurst, 1986; Schwartz & Sperling, 1983; Ullman, 1979), because a stimulus that has a perfect rigid interpretation is perceived as nonrigid. (It should be noted that the nonrigid interpretation also is a veridical 3D interpretation that is consistent with the 2D stimulus; it happens not to match the required response mapping.) These stimuli are multistable, yielding more than two possible stable percepts. In our experiments, when subjects perceived a nonrigid object, they were required to compute the name of one of the possible rigid objects that was consistent with what they perceived.

Relations to Previous Empirical Studies

We found that shape identification performance increases with the number of dots displayed and the extent of depth portrayed. Neither of these results is surprising. The numerosity result is an extension of previous, more subjective, measures of the depth perceived in simple KDE displays (Braunstein, 1962; Green, 1961). Increasing the number of dots provides the observer with more samples of the motion of the shape portrayed. Increasing depth extent increases the range of velocities used. Both manipulations increase the observer's signal-to-noise ratio in the task, in which noise sources may be both external (such as position quantization in the display and sparse shape sampling) and internal.

What is Computed in KDE?

Within measurement error, subjects performed equally well in the motion judgment task of Experiment 3 and comparable KDE tasks of Experiments 1 and 2. Further, the most common confusion error was the same in all experiments. And there is every reason to suppose that, if more data were available, the less common errors also would be highly correlated. In brief, we have succeeded in creating two equivalent tasks for classifying stimuli into 53 shape categories: One is solved by a KDE mechanism that yields a perceived 3D shape, and the other is solved by a motion perception mechanism that yields a perceived pattern of 2D motions. What does this imply about the mechanism of KDE and about the technology of KDE experimentation?

Although the specific nature of the perceptual algorithm that extracts 3D structure from 2D motion has not yet been established, it is reasonable to expect that it ultimately will be. Whatever the computation, the equivalent computation could, in principle, be carried out by some other system that was supplied with the same raw information, in this instance, the optical flow fields. In Experiment 3, we demonstrated that the measurements of the optic flow fields at six points provide sufficient information for the shape categorization task. When the optic flow at these locations is provided to observers in a response-compatible format, they can use this optic flow information to categorize the stimuli in perceived 2D just as efficiently as when they categorize KDE stimuli in perceived 3D. What is special about extracting structure from motion is not the informational capacity of the KDE system, but the perceptual capacity for extracting the relevant information and providing it perceptually as 3D depth.

For extracting structure from motion, the relevant information is optic flow. This was demonstrated in Experiment 2 (in which the residual nonflow cues were eliminated) and by experiments in which dots were given maximum lifetimes of only two (or three) frames so that correspondence cues were weakened and only optic flow cues survived (Dosher et al., in press; Landy et al., 1987). The relevant information in our particular shape discrimination task is the set of local velocity minima and maxima in the optic flow and their approximate shape. A reasonable assumption about the structure-from-motion computation is that the perceptual system automatically locates these maxima and minima, extracts the velocities, and transforms them into perceived depths. (Relative velocity has long been recognized as an extremely potent depth cue [e.g., Helmholtz, 1910/1924, p. 295ff; Rogers & Graham, 1979] and undoubtedly is a critical component of KDE.) When the relevant areas of optical flow are extracted instead by our display processor and presented to the subject as isolated patches, the subject is still able to classify the velocity in the patches, but the automatic perceptual conversion of velocity into perceived depth is inhibited. Nevertheless, the extracted velocity information is sufficient to enable accurate classification of the stimuli when a response-compatible format is made available.

Figure 6 illustrates the processes that are assumed to be involved in object recognition via the KDE. From the stimulus, the subject extracts a 2D velocity flow field. The KDE is the process whereby 3D depth values are extracted from the flow field. These depth values are combined with other shape and contour information from the stimulus to yield a 3D object percept which then forms the basis for the subject's response. A KDE-alternative computation is one that uses the same stimulus and velocity flow field, but circumvents the KDE computation by deriving the required response directly from the flow field. Experiment 3 demonstrated that a KDE-alternative computation would be possible in principle if the subject could extract the velocities at the six most relevant locations.

In transforming flow-field velocity into perceived depth, there is an inherent ambiguity in sign: A given velocity can equally well indicate depth toward or away from the observer. This ambiguity is inherent in the optics of the display and reflected in our scoring procedure. However, the perceptual system tends to resolve the ambiguity consistently in nearby locations. On those occasions in which it does not (e.g., when it interprets leftward motion as closer in one display area and as further in another), the display appears to be grossly nonrigid. The likelihood of consistent depth interpretation has been studied by Gillam (1972, 1976) and probably can be modeled by locally connected cooperative-competition networks (see Sperling, 1981, for an overview of cooperation–competition in binocular vision and Williams & Phillips, 1987, for an example of cooperation in motion perception).

KDE-Alternative Computations

It is useful to distinguish three kinds of computations: KDE, KDE-alternatives, and artifactual non-KDE computations. The KDE computation is an automatic perceptual computa-

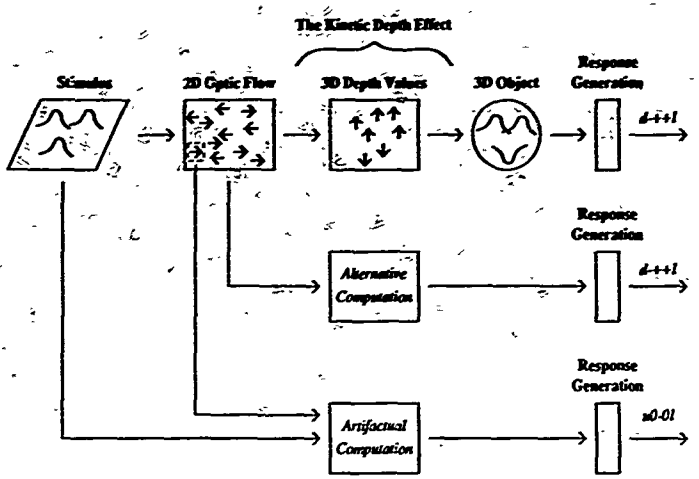


Figure 6. Flowchart for KDE, KDE-alternative, and artifactual computations. (From the stimulus, the following are assumed to be computed in sequence: 2D velocity flow field, 3D depth values [KDE computation], a 3D object representation [which in this instance happens not to correspond perfectly with the object represented by the stimulus], and the required response sequence. The KDE-alternative computation computes the required response sequence directly from the 2D optic flow without an intermediate stage of perceived 3D depth; that is, it simulates the KDE computation in another part of the brain. An artifactual computation uses incidental stimulus cues or motion cues from only a small part of the stimulus to arrive at a response.)

tion made, in the case of our stimuli, on velocity flow fields, and it results in perceived depth (a 3D percept) at those visual field locations where it is successful. A KDE-alternative computation is a computation on velocity flow fields similar to the KDE computation except that it is made consciously in some other part of the brain. It results in a knowledge of the correct response, but it does not yield perceived depth: The field is perceived as flat. An artifactual, non-KDE computation uses an incidental property of the display to compute the correct response, and the computation may be quite unrelated to the KDE computation. For example, the various objective studies of KDE that we considered in the beginning of this article all were vulnerable to computations that used only a small portion—in some instances only the movement of a single dot—of the stimulus information that would have been required by a KDE computation.

Of the five studies reviewed in the beginning of this article, the possible artifactual computations involved 1 dot (one study), 2 dots (two studies), and other cues (two studies). The problem is purely technical; the possible artifactual computations are quite different from KDE computations. There is a great risk of admitting an artifactual computation when the set of possible stimuli is small and when the required KDE computation itself is relatively simple. Even though subjects in these studies may have perceived KDE depth, a simple 2D strategy would have improved response accuracy. Although some of these procedures could have been improved, we deemed it better, from the outset, to use a large set of stimuli that can be identified only after a relatively elaborate KDE computation. What distinguishes the present task from prior tasks is that they admitted artifactual computations that were shortcuts to the correct response; the present alternative computation is an equivalent computation to KDE.

With respect to KDE-equivalent computations, we can ask two questions: Do they ever occur, and if they do, how can we be sure that they do not always occur? To demonstrate that a KDE-equivalent computation can occur we first have to know what the KDE computation itself is, and then to perturb the stimulus so that the automatic KDE computation cannot occur. In our experiment (and probably more generally), the essential KDE computation is the discovery of local velocity minima and maxima, and the consistent depth labeling of these minima and maxima. In Experiment 3, six stimulus areas around the velocity extrema were extracted from the KDE stimulus, and (in order to avoid the automatic KDE computation) they were presented as isolated squares. The subjects were able to label these areas consistently with respect to velocity (not depth, because the display was perceived as flat). Thus, subjects performed a KDE-equivalent task by means of a KDE-equivalent computation. Furthermore, the pattern of errors in the equivalent task corresponded to the previous error pattern in the KDE task. Although there are necessarily some differences between the KDE stimuli and the alternative stimuli, our strong result makes it clear

that, along with artifactual computations, the possibility of a KDE-alternative computation has to be considered in interpreting KDE experiments.

Artifactual computations are most easily discriminated from KDE computations by varying stimulus parameters. Stimulus cues that might support an artifactual computation are removed, masked or are rendered useless by irrelevant variation. If response accuracy survives, we have increased confidence that it is based on a KDE computation.

KDE and KDE-alternative computations use the same stimulus attributes; they differ in where in the brain the computation is made. Two tools for discriminating between these computations are *introspection* and *dual tasks*. For example, all subjects, without conscious effort, immediately perceive our KDE stimuli as solid 3D objects. When subjects honestly report that they perceived 3D depth in dynamic KDE stimuli, by definition, they have performed a KDE computation. The problem is that KDE may not be the only computation being performed. For complex stimuli such as ours, however, it is hard to imagine that a subject could be performing a useful alternative computation without awareness. Indeed, the discovery of an alternative computation for KDE is the structure-from-motion problem, and the solution proposed in Experiment 3 may be the first workable solution for stimuli of this type. It would be remarkable if subjects, even sophisticated subjects, discovered the solution in the course of viewing the stimuli. Still, even in this case, but especially with simpler stimuli, it would be better to use a formal procedure to exclude alternative computations. This requires, for example, (a) isolating the alternative computation, as in Experiment 3, (b) finding a concurrent task or similar manipulation that selectively interferes with the alternative computation relative to the direct KDE-computation, and (c) using the modified or dual tasks with the original stimuli.

An alternative KDE computation is analogous to an alternative stereoptic depth computation that is carried out by monocularly examining the left and right members of a stereogram. When stimuli are designed to take advantage of the exquisite sensitivity of stereopsis, an alternative monocular computation that uses remembered disparities is not feasible, even though it may be learnable in special cases. The same is undoubtedly true for KDE and alternative KDE computations: For complex KDE stimuli, viewed briefly, the alternative computation is simply out of the question. However, the problem of interpreting experimental results has not been alternative KDE computations but artifactual non-KDE computations. The best way to avoid subsequent problems of interpretation is to use complex stimuli, like the 53-shape stimulus set used here, that are matched to and challenge the ability of the human KDE computation.

Summary and Conclusion

A new shape identification task for measuring KDE performance is proposed. With its lexicon of 53 shapes, accurate identification requires either an accurate 3D shape percept or a KDE-alternative computation based on simultaneous measurements of 2D velocity in six positions of the display. Performance in the shape identification task improved with increased numerosity in a multidot display and with an increase in the amount of depth portrayed. Shape identification was not mediated by incidental texture-density cues but rather by motion cues derived from optic flow. The objective shape identification task is proposed as a sensitive measure of the critical aspect of kinetic depth performance. It is proposed that the structure-from-motion algorithm used by subjects to solve the KDE shape identification task involves finding local 2D velocity minima and maxima and assigning depth values to these locations in consistent proportion to their velocities.

References

Adelson, A. H. (1985). Rigid objects that appear highly non-rigid. *Investigative Ophthalmology and Visual Science, 26* (Suppl.), 56.

Andersen, G. J., & Braunstein, M. L. (1983). Dynamic occlusion in the perception of rotation in depth. *Perception & Psychophysics, 34*, 356–362.

Bennett, B. M., & Hoffman, D. D. (1985). The computation of structure from fixed-axis motion: Nonrigid structures. *Biological Cybernetics, 51*, 293–300.

Braddick, O. (1974). A short-range process in apparent motion. *Vision Research, 14*, 519–529.

Braunstein, M. L. (1962). Depth perception in rotating dot patterns: Effects of numerosity and perspective. *Journal of Experimental Psychology, 64*, 415–420.

Braunstein, M. L. (1977). Perceived direction of rotation of simulated three-dimensional patterns. *Perception & Psychophysics, 21*, 553–557.

Braunstein, M. L., & Andersen, G. J. (1981). Velocity gradients and relative depth perception. *Perception & Psychophysics, 29*, 145–155.

Braunstein, M. L., & Andersen, G. J. (1984). A counterexample to the rigidity assumption in the visual perception of structure from motion. *Perception, 13*, 213–217.

Clocksin, W. F. (1980). Perception of surface slant and edge labels from optical flow: A computational approach. *Perception, 9*, 253–269.

Dosher, B. A., Landy, M. S., & Sperling, G. (1989). Ratings of kinetic depth in multidot displays. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 816–825.

Dosher, B. A., Landy, M. S., & Sperling, G. (in press). Kinetic depth effect and optic flow: I. 3D shape from Fourier motion. *Vision Research*

Dosher, B. A., Sperling, G., & Wurst, S. A. (1986). Tradeoffs between stereopsis and proximity luminance covariance. *Vision Research, 26*, 973–990.

Gillam, B. (1972). Perceived common rotary motion of ambiguous stimuli as a criterion of perceptual grouping. *Perception & Psychophysics, 11*, 99–101.

Gillam, B. (1976). Grouping of multiple ambiguous contours. Towards an understanding of surface perception. *Perception, 5*, 203–209.

Green, B. F., Jr. (1961). Figure coherence in the kinetic depth effect. *Journal of Experimental Psychology, 62*, 272–282.

Helmholtz, H. v. (1924). *Helmholtz's Treatise on Physiological Optics* (J. P. C. Southall, Ed. and Trans.) Optical Society of America. Reprinted by Dover Publications, New York. (Original work published 1910)

Hildreth, E. C., & Grzywacz, N. M. (1986). The incremental recovery of structure from motion. Position vs. velocity based formulations *Proceedings of the Workshop on Motion Representation and Analysis, IEEE Computer Society #696, Charleston, South Carolina*

May 7–9, 1986 (pp. 137–144). Los Angeles: IEEE Computer Society Press.

Hoffman, D. D. (1982). Inferring local surface orientation from motion fields. *Journal of the Optical Society of America, 72,* 888–892.

Hoffman, D. D., & Bennett, B. M. (1985). Inferring the relative three-dimensional positions of two moving points. *Journal of the Optical Society of America A, 2,* 350–353.

Hoffman, D. D., & Flinchbaugh, B. E. (1982). The interpretation of biological motion. *Biological Cybernetics, 42,* 195–204.

Koenderink, J. J., & van Doorn, A. J. (1986). Depth and shape from differential perspective in the presence of bending deformations. *Journal of the Optical Society of America A, 3,* 242–249.

Landy, M. S. (1987). A parallel model of the kinetic depth effect using local computations. *Journal of the Optical Society of America A, 4,* 864–876.

Landy, M. S., Dosher, B. A., & Sperling, G. (1985). Assessing kinetic depth in multi-dot displays. *Bulletin of the Psychonomic Society, 19,* 23.

Landy, M. S., Sperling, G., Dosher, B. A., & Perkins, M. E. (1987). Structure from what kinds of motion? *Investigative Ophthalmology and Visual Science, 28*(Suppl), 233.

Lappin, J. S., Doner, J. F., & Kottas, B. L. (1980). Minimal conditions for the visual detection of structure and motion in three dimensions. *Science, 209,* 717–719.

Longuet-Higgins, H. C., & Prazdny, K. (1980). The interpretation of a moving retinal image. *Proceedings of the Royal Society of London, Series B, 208,* 385–397.

McKee, S. P., Silverman, G. H., & Nakayama, K. (1986). Precise velocity discrimination despite random variations in temporal frequency and contrast. *Vision Research, 26,* 609–619.

Petersik, J. T. (1979). Three-dimensional object constancy: Coherence of a simulated rotating sphere in noise. *Perception & Psychophysics, 25,* 328–335.

Petersik, J. T. (1980). The effects of spatial and temporal factors on the perception of stroboscopic rotation simulations. *Perception, 9,* 271–283.

Proffitt, D. R., Bertenthal, B. I., & Roberts, R. J. (1984). The role of occlusion in reducing multistability in moving point-light displays. *Perception & Psychophysics, 36,* 315–323.

Rogers, B., & Graham, M. (1979). Motion parallax as an independent cue for depth perception. *Perception, 8,* 125–134.

Schwartz, B. J., & Sperling, G. (1983). Luminance controls the perceived 3-D structure of dynamic 2-D displays. *Bulletin of the Psychonomic Society, 21,* 456–458.

Sperling, G. (1981). Mathematical models of binocular vision. In S. Grossberg (Ed.), *Mathematical psychology and psychophysiology.* Providence, Rhode Island: *Society of Industrial and Applied Mathematics–American Mathematical Association (SIAM–AMS) Proceedings, 13,* 281–300.

Todd, J. T. (1982). Visual information about rigid and nonrigid motion: A geometric analysis. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 238–252.

Todd, J. T. (1984). The perception of three-dimensional structure from rigid and nonrigid motion. *Perception & Psychophysics, 36,* 97–103.

Todd, J. T. (1985). The analysis of three-dimensional structure from moving images. In D. Ingle, M. Jeannerod, & D. Lee (Eds.), *Brain mechanisms and spatial vision* (pp. 73–93). The Hague, The Netherlands: Martinus Nijhoff.

Ullman, S. (1979). *The interpretation of visual motion.* Cambridge, MA: MIT Press.

Ullman, S. (1984). Maximizing rigidity: The incremental recovery of 3-D structure from rigid and non-rigid motion. *Perception, 13,* 255–274.

Wallach, H., & O'Connell, D. N. (1953). The kinetic depth effect. *Journal of Experimental Psychology, 45,* 205–217.

Webb, J. A., & Aggarwal, J. K. (1981). Visually interpreting the motion of objects in space. *Computer, 14,* 40–49.

Williams, D., & Phillips, G. (1987). Cooperative phenomena in the perception of motion direction. *Journal of the Optical Society of America A, 4,* 878–885.

Received July 8, 1988
Revision received October 13, 1988
Accepted October 14, 1988 ■

Correction to Driver and Baylis

In the article, "Movement and Visual Attention: The Spotlight Metaphor Breaks Down," by Jon Driver and Gordon C. Baylis (*Journal of Experimental Psychology: Human Perception and Performance,* 1989, Vol. 15, No. 3, 448–456), the display durations were incorrect and should be doubled to give the correct figures. Each display frame actually lasted 40 ms. Thus, total display duration was 200 ms in Experiments 1, 3, and 4 and was 120 ms in Experiment 2.

Journal of Experimental Psychology:
Human Perception and Performance
1989, Vol. 15, No. 4, 816-825

0096-1523/89/$00.75

Ratings of Kinetic Depth in Multidot Displays

Barbara A. Dosher
Columbia University

Michael S. Landy and George Sperling
New York University

Subjects saw kinetic depth displays whose shape (sphere or cylinder) was defined by luminous dots distributed randomly on the surface or in the volume of the object. Subjects rated perceived 3-D depth, rigidity, and coherence. Despite individual differences, all 3 ratings increased with the number of dots. Dots in the volume yielded ratings equal to or greater than surface dots. Each rating varied with 3 of 4 factors (shape, distribution, numerosity, and perspective), but the ratings either between trials or between conditions were often uncorrelated. Object shape affected rigidity but not depth ratings. Veridically perceived polar displays had slightly lower rigidity but higher depth ratings than parallel projection displays. (Reversed polar displays were always grossly nonrigid.) The interaction of ratings and stimulus parameters requires theories and experiments in which different KDE ratings are not treated interchangeably.

When a two-dimensional (2-D) projected image corresponds to a three-dimensional (3-D) object that is rotating, viewers frequently perceive an object with depth. Because rotation induces apparent 3-D depth even when isolated still views of the object fail to induce perceived depth, the phenomenon is called the *kinetic depth effect* or KDE (Wallach & O'Connell, 1953). In this article, experiments are discussed that consider the perception of dot displays, in which each stimulus consists of illuminated dots on an otherwise invisible object. It will be demonstrated that there are a number of partially decoupled aspects to the perception of these displays under motion. (a) Coherence, whether all dots in the display are seen as constituting a single object, (b) depth, the amount of 3-D depth seen in the display, (c) rigidity, whether those illuminated dots that are perceived as constituting a coherent object also are perceived as maintaining their relative 3-D positions (rigid appearance) or as changing their relative positions (nonrigid, rubbery appearance).

There is a large body of literature examining the function of various kinds of stimulus variables in the kinetic depth effect. Some of the classic stimulus variables include the number of elements defining the stimulus, element shape, occlusion, perspective, correspondence, element density, and rotation speed. The effects of these stimulus variables were examined by a variety of dependent measures, global "goodness" judgments (Andersen & Braunstein, 1983, Braunstein, 1962; Braunstein & Andersen, 1984, Green, 1961, Petersik, 1980), qualitative motion categorization (surface, rotary, oscillatory; Caelli, 1979, 1980), judgments about objective rotation direction (Braunstein, 1962, 1977, Petersik, 1979, 1980), perceived curvature or shape (Braunstein & Andersen, 1984, Todd, 1984), and proportion of corresponding elements across frames (Lappin, Doner, & Kottas, 1980). One question that arises is whether the choice of dependent measure is of no consequence. Are all these measures simply reflections of a unitary aspect of the kinetic depth percept? In order to answer this question, we examined the independence of the three aspects of the percept listed above by collecting three separate responses on every trial in experiments that varied some important stimulus variables.

As a concrete example, consider an early set of experiments by Green (1961). He examined, among other factors, the importance of the number of stimulus elements on the KDE. Subjects were asked to rate displays on a scale that combined the notions of rigidity and coherence, as defined here. The label goodness is used here to describe Green's combined rating scale, in order to distinguish it from our use of the distinct labels rigidity and coherence (which Green used interchangeably). Green demonstrated that the number of stimulus elements was a potent factor in determining the goodness of a perceived object under various forms of rotation, generally, the more stimulus elements, the higher the rated goodness, with the largest increments occurring with the number of elements under 32. In principle, the increment in goodness could have reflected some unspecified weighting of coherence and rigidity. It is not clear whether numerosity affected one or both of these aspects of the percept primarily, nor is it clear how it affected perceived depth of element trajectories.

Here we investigated element numerosity, as well as a number of other factors that may vary in viewing 2-D projected images of objects. In particular, we examined one image projection factor (parallel projection versus perspective projection, with projection distance at three times object radius) and three object factors (the number of elements representing an object, from 4 to 80 elements; whether the elements representing the object were entirely on the surface or distributed throughout the volume, and the strength of density cues

The work described in this article was carried out in the Psychology Department of New York University and supported by the Office of Naval Research, Grant N00014-85-K-0077 and the USAF Life Sciences Directorate, Visual Information Processing Program Grants 85-0364 and 88-0140.

Correspondence concerning this article should be addressed to Barbara A. Dosher, Psychology Department, Columbia University, Box 28, Schermerhorn Hall, New York, New York 10027.

Atch- 18

to depth in a still frame, by using different forms). Our aim was to determine the effect of these KDE stimulus variables on ratings of coherence, depth, and rigidity.

Our results corroborate some findings of previous investigators, for example, that the number of dots in the object and the presence of polar perspective can add to the strength of KDE. However, we also show that these stimulus variables do not generally affect all three aspects of the KDE percept equally and that there are many subtleties and complexities in the KDE.

Method

Because of the large number of stimulus variables, the study was divided into three separate experiments. The experiments were conducted with the same subjects and with the same procedures, except as noted.

Subjects

There were 4 subjects in the experiments, including 2 of the authors of this article and 2 students. The students were paid for their participation. Three subjects had normal or corrected-to-normal vision; subject CFS could be corrected only to 20 40.

Apparatus

All stimuli were computer generated, and the display and response collection was computer controlled. Experiment 1 and a pilot experiment used a point/vector display controller (Kropfl, 1975) and an HP1304A display monitor. Display resolution was 1024 × 1024 pixels. Experiments 2 and 3 used a raster display controller, Adage RDS-3000, and a Conrac 7211C19 RGB color monitor. Display resolution was 512 × 512 pixels. Experiment 1 used binocular viewing in a completely dark room. In Experiments 2 and 3, subjects viewed the stimuli monocularly through a reduction tube, with an aperture slightly larger than the stimuli. Hence, weak stereo cues to flatness may have been present in Experiment 1, but not in Experiments 2 and 3.

Stimuli

Stimuli consisted of random white dots scattered on the surface or throughout the volume of invisible spheres and cylinders. The probability distribution used for dot placement was uniform across the surface (or through the volume) in each case, but choices of dots were constrained so as to fill the surface or volume fairly evenly by partitioning into equal-area (or equal-volume) segments and putting equal numbers of dots in each segment. Five stimulus parameters were varied. First, there were two types of objects, a sphere of diameter 2° of visual angle and an upright cylinder of height 2° of visual angle and cross-sectional diameter 2° of visual angle. The number of dots was varied from 4 to 80. These dots were either positioned on the surface or in the volume of the object being simulated. Stimuli were either presented in parallel projection (i.e., with no perspective) or with an exaggerated amount of polar perspective (corresponding to a viewing distance of three times the object radius, far smaller than the actual viewing distance). All stimuli were rotated about a vertical axis through the center of the simulated object. Stimuli were either rotating front-left or front-right, although this distinction is only meaningful for the stimuli with polar perspective. Single-frame views of some sample stimuli are shown in Figure 1.

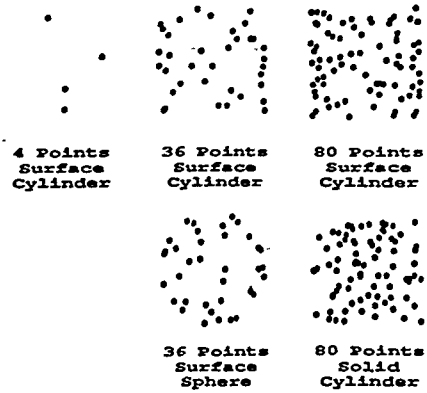


Figure 1. Single frames from some sample stimuli varying in numerosity, distribution, and form.

Procedure

On each trial, the subject was shown a fixation target, which then disappeared and was followed shortly by one rotation of the stimulus (See Table 1 for details of rotation speeds, etc.) After the stimulus presentation was complete (approximately 2 s), four responses were required of the subject. First, the subject indicated the direction of rotation of the object (front-left or front-right). These responses were used in polar projection displays to determine whether the subject perceived the object in the veridical or the reversed mode. Then, three different ratings of the percept were required: depth, coherence, and rigidity.

Depth rating. The subject indicated the amount of depth perceived in the stimulus on a scale from 1 to 5. Given that all stimuli were based on objects rotating about a vertical axis, depth was related to an inferred "top view" of the stimulus. The subject was shown the top views (Figure 2) to facilitate his or her rating. The most depth, 5, was associated with a perceived circular path for each dot; the least depth, 1, was associated with no perceived depth and hence an oscillatory linear path for each dot.

Coherence rating. The next rating, also on a scale of 1 to 5, was of the perceived coherence of the multidot display. A rating of 5 indicated the greatest coherence (i.e., all the dots held together as one object). A rating of 4 indicated that a few dots did not cohere, 3 indicated that the display broke up into two distinct objects (segmentation), 2 indicated that three or more objects were perceived, 1 indicated there was no perceived coherence whatsoever.

Rigidity rating. Perceived rigidity was rated on a scale from 1 to 5, with a 5 indicating one or more totally rigid objects, and lower numbers indicating more and more nonrigidity or "rubberness."

Figure 2. The inferred top views of the stimuli that were used to define the five levels of perceived depth ratings.

Subjects judged all three aspects of each percept. This allowed us to relate the three aspects on an individual trial basis. Had the three judgments been collected separately, the relation between the three judgments would have been available only at the level of the mean data.

Designs

Experiments 1, 2, and 3 differed in the factors varied and in the display devices used. Table 1 summarizes the design and viewing conditions for each experiment. All designs were fully crossed in each included factor. Left and right veridical rotation direction was also a factor in each experiment. Stimulus order was randomized within block and each block consisted of one token of each stimulus type, yielding 64 stimuli per block in Experiment 1 and 32 stimuli per block in Experiments 2 and 3. A different random token of each stimulus type was generated for each of six blocks per experiment. A pilot experiment yielded no effect on any response ratings of object size (2° of visual angle vs. 4° of visual angle), 2° of visual angle was used subsequently. The polar perspective manipulation was defined with respect to object radii, so that the object size manipulation also varied the mismatch between actual and appropriate viewing distance for the degree of perspective. This and some of the current work was originally reported in Landy, Dosher, and Sperling (1986).

Results

The results for Experiments 1 and 2, pooled across subjects, are shown in Figure 3. There were significant individual differences, discussed below, and so statistical analyses were performed as within-subject analyses of variance (ANOVAs). The 12 trial-type replications that resulted from collapsing over rotation direction and test block formed the random factor. These replications represented responses to 12 distinct, randomly generated stimuli of each type. Table 2 lists the significance levels associated with each rating, factor, subject, and experiment, along with a qualitative coding of the direction of the results. Table 2 thus gives a quick summary of the consistency both between subjects and within subjects across experiments. Table 3 summarizes the results of previous related experiments. Notice that the current set of experiments include factors and ratings that are either unrepresented in the literature or represented by a questionable combined measure.

Numerosity. In Experiment 1, in which the number of dots ranged from very small to moderate in number, all three ratings for 3 of the 4 subjects were increased by increasing the number of dots. The 4th subject showed a very different behavior (e.g., see Figure 4). The four-dot stimuli yielded very high depth ratings for this subject. The subject mentioned afterward that the stimuli reminded him of organic chemistry drawings and yielded a vivid percept.

In Experiment 2, in which the number of dots was moderate to large, the effect of the number of dots was less dramatic. Depth ratings increased slightly and saturated at these high numerosities. At these larger levels of numerosity, the effects of numerosity on coherence and rigidity were small. There were no significant effects on coherence ratings and numerosity was related to rigidity ratings for only 2 of 4 subjects. In summary, by increasing dot numerosity, all three ratings increased, up to a point, and then saturated. Depth ratings appeared to increase and saturate in a continuous manner, whereas coherence and rigidity ratings were high for all but the sparse displays (eight or fewer elements). These findings are in general agreement with those of Green (1961) over similar ranges of numerosity. However, Green's judgment was one of overall goodness and more nearly agrees with the depth judgments reported here.

Intensity. In Experiment 3, in which the intensity of displays was increased from 0.86 to 42.7 μcd/dot, there was no significant effect on ratings (with the exception of a single subject on a single rating). We ruled out an effect of varying display types in which overall stimulus intensity (contrast) varies in the visible range. (But see Dosher, Landy, & Sperling, in press, for manipulations of intensity very near to threshold, which do affect kinetic depth performance.)

Form. For conditions in which a spherical shape was directly contrasted with a cylinder (Experiment 1), the sphere was rated more rigid than the cylinder by all subjects and more coherent by 3 of the 4 subjects. The higher rigidity ratings for spheres overall was actually due to a strong interaction between form and perspective. Rigidity ratings were differentially lower for cylinders under perspective. The sphere gives less representation to dots that are substantially affected by the projection factor (far from the axis of rotation), and the increase in perceived nonrigidity may have resulted

Table 1
Experimental Factors and Conditions

Experiment	Numerosity	Form (cylinder or sphere)	Distribution (surface or volume)	Perspective (parallel or polar)	Luminance
1	4, 8, 16, 36	both	yes	yes	1.45 μcd/dot[a]
2	36, 48, 64, 80	cylinders	yes	yes	3.02 μcd/dot[b]
3	48	cylinders	yes	yes	0.86, 3.02, 11.52, 42.72 μcd/dot[b]

[a] Point plot display, resolution 1024 × 1024 pixels; 36 new frames per 360° rotation (or 10° per frame), 60 ms per new frame, or 2.16 s per full rotation; dark room; binocular free viewing; viewing distance 1.1 m; object diameters 2° visual angle (parallel perspective).
[b] Raster display, resolution 512 × 512 pixels; 36 new frames per 360° rotation (or 10° per frame); 66.67 ms per new frame, or 2.4 s per full rotation; dim room (8 cd/m²) with light-tight viewing hood; monocular viewing through a reduction aperture; viewing distance 1.6 m; object diameters 2° visual angle (parallel perspective).

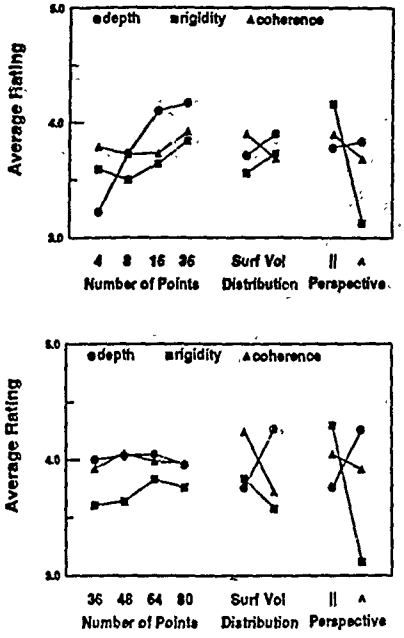


Figure 3 Response data for all ratings and stimulus manipulations in Experiments 1 (upper panel) and 2 (lower panel), pooled across subjects. (The parameter is the particular rating made: perceived depth [circles], rigidity [squares], and coherence [triangles]. In each panel, the first set of curves is for the number of dots in the stimulus, the second for the effect of distributing the dots across the surface or throughout the volume of the object, and the third for the effect of perspective transformation [parallel or polar].)

in object breakdown (segmentation, incoherence) in some cases. Braunstein and Andersen (1984) also compared spheres to cylinders. They used each form as the base for elliptical distortions and found that sensitivity to minor axis variation (flatness of the elliptical orbits) was slightly greater when the base form was a cylinder than when the base form was a sphere. However, this was a cross-experiment comparison with different groups of subjects in the different conditions.

Distribution. The effect of dot distribution (in the volume or on the surface) was generally small with significant individual variation (see Figures 3 and 4 and Table 2). Close examination of Table 2 suggests that distribution was more important when numerosity was large and when there was reduced single-view shape information (i.e., for the cylinder, Experiments 2 and 3). In these conditions, distribution in the volume increased depth ratings for 3 of the 4 subjects and might have improved coherence for some subjects as well.

Green's (1961) overall goodness measure showed slightly higher scores for surface representations than for completely random placements in the volume of a cube, but an enormous benefit for point representations with regular placements in the volume. Our random sampling incorporated partition equality and hence represented a compromise between Green's random and regular conditions. Dots in the volume might have increased depth ratings because a range of intermediate velocities was represented in the cylinders, whereas in surface representations of cylinders, all dots were traveling at more nearly the same velocity, except at the edges of the object. To the extent that differential velocity supported depth segregation (Braunstein & Andersen, 1981), representation of intermediate velocities may have been useful. The ability of distribution to strongly affect the kinetic depth percept may have also depended on the unavailability of other strong cues to shape such as perspective, texture density, or contour (see the discussion of Figure 5 below).

Perspective. For all subjects, the rigidity ratings were decreased by adding polar perspective. The effect of perspective on coherence was small and depended on the subject. A collateral analysis of the polar perspective trials that sorted those occasions on which the perceived rotation direction disagreed with the intended rotation direction found that most, but not all, of the decrease in rated rigidity with polar projection occurred when the observer perceived the stimulus in the reversed mode (see also Gregory, 1970; Schwartz & Sperling, 1983). Thus, when polar displays were perceived in their reversed mode, they appeared grossly nonrigid; when polar displays were perceived veridically, they appeared slightly less rigid than the corresponding parallel displays.

Neither our polar stimuli nor our parallel stimuli were viewed at the appropriate viewing distance. The parallel stimuli would have to be viewed from infinity; the polar stimuli from 6 cm; the actual viewing distances were in the range 1 m in the various experiments. (Had we produced appropriately projected objects for the 1 m viewing distance, they would have been negligibly different from the actual parallel stimuli. When viewed at the appropriate viewing distance of 6 cm, our polar displays possessed little depth—largely a consequence of the large scale.) The greater mismatch between appropriate viewing distance and the actual viewing distance for polar stimuli conceivably might have accounted for the fact that veridically perceived polar stimuli received slightly lower rigidity ratings than parallel stimuli. But this distance mismatch does not bear on the overwhelming cause of nonrigidity in polar displays—that stimuli are perceived in reversed mode. Even the secondary effect of polar projection on rated rigidity may have depended only weakly on projection/viewing distance mismatch. As noted previously, a pilot study in which object size was varied by a factor of 2:1 (producing a change between projected and actual viewing distance of 2:1) had no significant effect on any rating. Finally, Cutting (1987) found little impact of mismatch between simulated and actual viewing distances.

The rigidity and coherence results reported here agree with the reported relationship between the amount of perspective and the ability to infer the intended rotation direction (Braun-

Table 2
Significant Factors in Experiments

Experiment/ Subjects	Numerosity	Form	Distribution	Perspective	D × P
			Depth judgment		
1	(small numbers)				
MSL	+,•••	ns	+,•••	ns	••
BAR	+,•••	ns	+,•••	+,••	•
CFS	+,•••	ns	+, ns	ns	ns
RHS	U,•••	+,••	−,••	ns	ns
2	(large numbers)				
MSL	~,••	—	+,•••	+,•••	•••
BAR	~,••	—	+,•••	+,••	•
CFS	~,•	—	+,••	ns	†
RHS	~,•••	—	−,•••	ns	ns
3					
MSL	—	—	+,••	+,•••	†
BAR	—	—	+,•	+,•	•
CFS	—	—	+, ns	+,•	†
RHS	—	—	−,•	ns	ns
			Coherence judgment		
1	(small numbers)				
MSL	+,•••	+,•	ns	−,•••	•••
BAR	+,•••	ns	+,••	+,••	•
CFS	+,•••	+,••	ns	−,•••	ns
RHS	+,•••	+,••	−,•••	−,••	ns
2	(large numbers)				
MSL	ns	—	ns	ns	ns
BAR	ns	—	+,•••	+,•••	•••
CFS	ns	—	−,••	−,•••	ns
RHS	ns	—	−,•••	−,•••	ns
3					
MSL	—	—	+,•	ns	†
BAR	—	—	+,••	+,•	••
CFS	—	—	+,•	−,•	ns
RHS	—	—	−,••	−,•	•
			Rigidity judgment		
1	(small numbers)				
MSL	+,•	+,••	+,†	−,•••	••
BAR	+,•••	+,••	+,••	−,•••	ns
CFS	+,••	+,••	+,†	−,•••	ns
RHS	X,•••	+,•	+,•	−,•••	•
2	(large numbers)				
MSL	ns	—	ns	−,•••	ns
BAR	~,••	—	+,••	−,•••	ns
CFS	~,•	—	−,••	−,•••	ns
RHS	ns	—	−,•••	−,•••	ns
3					
MSL	—	—	ns	−,•••	ns
BAR	—	—	+,•	−,•••	ns
CFS	—	—	ns	−,•••	ns
RHS	—	—	ns	−,••	ns

Note. The p values (see below) are the significance of corresponding F values from an analysis of variance for each subject treating rotation direction and tokens of stimuli as the random factor. The symbols ~, +, and − indicate the pattern of the effect and can be referenced to the legend list below for each factor. See the text for a discussion of the interaction of distribution and perspective. Numerosity: +, increasing with number of points; ~, saturates with large number of points; U, U-shaped function of number of points. X, highest for smallest and largest number of points. Form: +, sphere > cylinder. Distribution: +, volume > surface. Perspective: +, polar > parallel. D × P = interaction of distribution and perspective. ns = not significant. Dashes = not applicable.
† $p < .100$. • $p < .05$. •• $p < .01$. ••• $p < .001$.

Table 3
Summary of Related Results

Judgment type	Numerosity	Form (cylinder or sphere)	Distribution (surface or volume)	Perspective (parallel or polar)
Depth	=Petersik, 1980			+Braunstein, 1962 =Petersik, 1980
Coherence Rigidity	−Braunstein, 1962 =Petersik, 1980			−Petersik, 1979 −Braunstein, 1977
Combined	+~Green, 1961	−Braunstein & Andersen, 1984	−Green, 1961	−Braunstein, 1962 =Braunstein, 1977

Note. The symbols ~, +, and − indicate the pattern of the effect and can be referenced to the note for Table 2. The symbol = indicates no effect. A summary of some relevant factors in prior experiments follows: Braunstein (1962), dots in volume of cube, $N = 2$–6, depth and coherence/rigidity judgments; Braunstein (1977), dots in volume of sphere, $N = 1{,}000$, varied perspective in horizontal and vertical dimensions, direction and coherence/rigidity judgments; Braunstein & Andersen (1984), dots on surface of sphere or ellipses, $N = 140$–160, shape and quality judgment; Green (1961), dots or line elements in volume or on surface of cube, $N = 4$–64, goodness rating (combined segmentation and rigidity); Petersik (1979), dots in volume of sphere, $N = 4$–45, depth and direction rating; Petersik (1980), dots in volume of sphere, $N = 5$–60, depth and direction rating.

stein, 1977; Petersik, 1979). The difference in the effect of perspective on rated rigidity and on rated coherence may explain the inconsistent results of Braunstein, who found that perspective decreased a combined rating of coherence and rigidity in one study (1962), but had no effect in another (1977).

Perspective generally increased the rated depth (shape) of the percept (although not all contrasts were significant, see Table 2). A prior study by Braunstein (1962, see Table 3) also found that perspective improved a "strength of depth" rating. Petersik (1980) found that depth judgments were not affected by perspective. However, this same study found no effect of numerosity ($N = 5$–60) on depth, which suggests that the experiment had insufficient power.

Interaction of Perspective and Distribution. By adding polar perspective, the rated depth was increased. Distributing dots throughout the volume of the object had the same effect. However, when the two were combined, a further increase was not achieved. This interaction between perspective and dot distribution is illustrated in Figure 5, and the significance levels for each subject are listed in Table 2. (Here again, the effect of distribution was greater in high-numerosity cylinders, Experiments 2 and 3.) As suggested above, some factors such as distribution may be more likely to control the percept in the absence of other strong cues to shape.

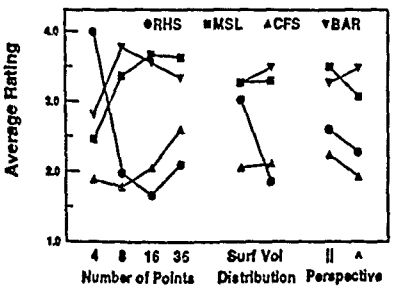


Figure 4 Coherence ratings in Experiment 1 (The parameter is the particular subject. Note the large individual differences.)

A Large Individual Difference. Occasionally, individual differences were very striking. An example of this is shown in Figure 4. Here the coherence ratings for all conditions of Experiment 2 are shown for individual subjects. Subject RHS was the only subject for whom the increased number and distribution of dots in the volume of the object decreased the coherence of the kinetic depth percept. Individual differences presumably occurred in earlier studies, but were undetected because prior studies collected few observations from each subject and performed cross-subject analyses. For another example of large individual differences in KDE, see Dosher, Sperling, and Wurst (1986).

Three Ratings Are More Informative Than One. So far, we have described the empirical results with respect to manipulations of dot numerosity, perspective, and so forth. What was perhaps most important was the added information gained by having multiple ratings of the stimuli. These ratings, each of which can be (and has been in the literature) construed as a measure of the "strength" of a KDE percept, did not necessarily covary. In many cases, as we have seen, an experimental manipulation had a different effect on different ratings. For example, shape significantly affected mean ratings of rigidity, but had little or no effect on depth ratings. At high numerosity, further increases in numerosity continued to increase depth ratings, but did not affect coherence ratings, and so forth.

Correlations Between Ratings. It was also possible to do a finer-grained analysis of the three ratings that looked beyond the means to the correlations on a trial-by-trial basis. Table 4 gives the trial-by-trial correlations pooled over conditions and subjects. Seven of the nine correlations in Table 4 are between −0.12 and +0.19; and the two highest correlations are still

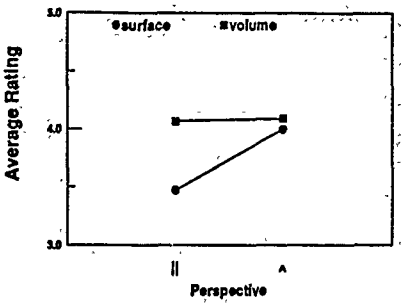


Figure 5. Interaction between perspective transformation and dot distribution in Experiments 1 and 2 (pooled).

rather low, with each relationship accounting for only about 12% of the variance. Although the subjects were affected somewhat differently by some of the experimental factors, on the whole, different subjects tended to use the ratings quite similarly. Such low correlations between ratings clearly rule out a simple, single factor interpretation that would require high positive or high negative correlations between each pair of ratings. For every subject, at least two of the three interrating correlations are low. Obviously, the rated qualities of kinetic depth percept reflect at least two underlying dimensions. Although the experiments were not designed in a way that would expose the KDE depth percept to multidimensional analysis, they were sufficient to bring this inherent multidimensionality to the fore. The fact that different ratings weigh differently on these dimensions cannot continue to be overlooked in KDE research.

Discussion

Wide Range of Percepts. The KDE for a multidot display is quite rich. When viewing a stimulus with a small number of dots, there generally are many possible stable percepts. Even though the whole was geometrically derived from a rigid object, perceptually, subgroups of dots form clusters, and each subgroup appears to move independently in 3-D, acting as a separate object. Groups of two or three dots can be perceived as moving independently in the plane, or as a 3-D and rigid configuration, or as a nonrigid 3-D configuration similar to the Ames window (as in Gillam, 1975, 1976, in which line segments were used rather than dots). In short, groups of dots do not necessarily cohere as single objects, even when they are being perceived in depth and when a unitary rigid 3-D interpretation is available. Even when dots are correctly perceived as parts of one coherent object, there is a range of possible percepts that differ in shape, in depth, and in perceived rigidity. The perceived coherence, shape, and rigidity have a complex and partially decoupled relationship.

Table 4
Correlations Between Judgments: All Subjects

Judgment types	Experiment 1	2	3
Depth–rigidity	.05	−.12	−.01
Depth–coherence	.08	.07	.16
Rigidity–coherence	.36	.35	19

Decoupled Aspects of Percepts. Some degree of decoupling between aspects of a KDE percept has been known since Braunstein (1962) varied perspective in KDE displays and observed an inverse effect on mean judgments of depth and of combined rigidity/coherence. However, it has implicitly been assumed that the manipulation of rigidity by perspective was a special case. The current experiments demonstrate that this decoupling between the various aspects of the percept is not restricted to the independent variation of perceived rigidity but is quite general because different factors affect different judgments. In terms of mean ratings in our experiments, rated depth was significantly affected by numerosity, distribution, and perspective. Rated segmentation was affected primarily by numerosity; this effect reflected a division between sparse and dense levels of numerosity (above or below 16 elements). Secondarily it was affected by form and perspective. Rated rigidity was primarily affected by perspective and numerosity. Additionally, correlations among the three ratings were low and sometimes negative when measured on a trial-by-trial basis.

Experiments in the literature on multidot KDE have used as the dependent measure either ratings or paired comparisons on some judgment dimension (see Table 3). The judgment dimensions either selected from a variant of the three ratings used here or combined two or more in one rating. Conflation of the dependent measure may, in part, explain some of the inconsistencies in the literature noted above. In particular, the combined coherence and rigidity ratings of Braunstein (1962, 1977) may account for the inconsistent effect of perspective in those studies.

Importance of Independent Factors. Our experiment manipulated a number of factors within a subject, factors that previously had been examined in separate experiments or had been chosen arbitrarily as fixed factors that happened to differ, along with the dependent measure, between studies in the literature. Shape, distribution, and numerosity have usually varied haphazardly between experiments. For example, Braunstein (1962, 1977) found inconsistent patterns of perspective on a rigidity–coherence judgment using 2–6 dots in the volume of a cube and 1,000 dots in the volume of a sphere, respectively. Braunstein (1962) and Petersik (1980) found inconsistent patterns of perspective on depth judgments using 2–6 dots in the volume of a cube and 4–45 dots in the volume of a sphere, respectively. It has been difficult to know whether the structural and numerosity factors explained the inconsistency in patterns.

The results of our experiments can be viewed as filling in Table 3 with a self-consistent set of data and providing previously unavailable data in the empty cells. The results clearly separate three important aspects of a kinetic depth percept: depth (shape), coherence, and rigidity. Because our stimulus parameters are manipulated within subject, cells are directly

comparable. A number of our results are similar to Green (1961), Braunstein (1962, 1977), and others. In other cases, in which inconsistent findings were reported, we suggested explanations based on confounded dependent measures. Typically, inconsistent results between KDE experiments result from judgments that combine component aspects (e.g., depth, coherence, rigidity) in unspecified, but probably different weightings.

Nonmotion Cues to Depth in KDE Displays. There are two classes of cues to object structure in our displays: static cues such as density and 2-D object contour and dynamic or motion cues that depend either on optic flow or on changing interpoint distances. Based on our data, the greatest likelihood of perceiving the veridical shape occurs with perspective images of spheres (rather than cylinders), with high dot numerosity, and with dots in the volume (rather than on the surface).

High dot numerosity guarantees a good representation of the 2-D contour of the sphere and of 2-D density cues. Even when the 2-D contour is not as suggestive, as in the case of the cylinder, the density cues that become visible with high numerosity may be important in providing static cues to shape. (In the case of surface distribution of elements, the 2-D density will increase toward the edges, whereas in volume distribution of elements, the density cue is reversed, as in Figure 1.) The presence of 2-D cues to shape, whether from contour or density, may constrain the perception. High element numerosity also minimizes the likelihood of atypical clumping or grouping characteristics likely in low numerosity figures, which then are likely to cause grouped or segmented (i.e., incoherent) percepts.

Perspective may simply serve as an additional cue to depth organization. Alternatively, the exaggerated perspective used here may be especially effective because it slightly increases the proportion of elements moving in the same direction, yielding a display similar to that arising from an image with occlusion, and possibly allowing stronger input to an optic flow analysis at high numerosity (J. Todd, personal communication, March 1987).

Distribution in the volume provides a range of velocities in any local area and may support a full depth percept by relating distance from the axis of rotation to dot velocity. Dot fields of different velocity, whether adjacent or superimposed, tend to segregate in depth (Braunstein & Andersen, 1981).

Percept Description Versus Objective Task Measures. An alternative to the measurement of one or another aspect of the kinetic depth percept by rating is to conceptualize a different sort of question. In rating, we ask about various aspects of the percept itself. An alternative is to ask whether a percept, whatever its subjective appearance, is adequate to support objective performance on a particular kind of judgment, such as a judgment of shape. Some attempts have been made in this regard. For example, Todd (1984) required subjects to make objective curvature judgments under various levels of nonrigidity in the kinetic depth image. Lappin et al. (1980) required subjects to make objective, paired-comparison judgments of the degree of correspondence in two-frame displays. We investigated one possible objective measure of having perceived shape from a kinetic depth display (Dosher, Landy, & Sperling, in press; Landy, Sperling, Dosher, & Perkins, 1987; Sperling, Landy, Dosher, & Perkins, 1989) This objective measure requires subjects to identify the object perceived from among a large lexicon of possible objects and offers an attractive alternative to the elaboration of subjective methods under study here.

Relation to Models. Three classes of computational models have been proposed to account for the kinetic depth effect based on motion cues: those deriving shape from optic flow fields (Clocksin, 1980; Koenderink & van Doorn, 1986), those deriving analytic solutions by assuming rigidity from m views of n points (Hoffman & Bennett, 1985; Ullman, 1979), and those based on maximizing rigidity in interpoint distances (Hildreth & Grzywacz, 1986; Landy, 1987; Ullman, 1979, 1984). Usually, flow-field models are applied to objects composed of densely packed points, and interpoint-distance models are applied to images composed of less than a few dozen points. Interpoint-distance models apply geometric computations to the 2-D image-plane positions of the given points to compute a 3-D object that is either totally rigid (Ullman, 1979) or a 3-D object that deforms minimally between adjacent frames (Landy, 1987; Ullman, 1984).

We will illustrate the problems that rigidity models have with data such as ours by considering, as an example, the incremental rigidity algorithm of Ullman (1984). When an n-point 3-D object undergoes rotation, the algorithm takes as its input a sequence of frames that represent the 2-D image-plane x, y projections of the n points. For each frame, the algorithm outputs an estimated depth value z for each point plus one overall fidelity score. The computation consists of a gradient descent in the space of depth values z to maximize the fidelity criterion. This criterion measures the deformation (nonrigidity) in the recovered 3-D object between the current frame and the prior frame.

To evaluate such an algorithm as a psychological model, one must associate quantities produced by the algorithm with aspects of human perception. We have shown here three aspects of performance that are partially separable in performance measures: segmentation, depth, and rigidity. Consider what happens in the case of four-point displays. Ullman's (1984) algorithm, like most others, simply assumes element correspondence and figural segmentation as prior processes For four-point objects, the incremental rigidity algorithm would have recovered the veridical single object with rigid depth assignments for all nonperspective images in the experiment. On the other hand, only 1 of our 4 subjects regularly perceived four-point displays as unitary; for the other subjects, these were usually perceived as two or more objects moving independently. This grossly violates the segmentation assumed by the Ullman model.

The algorithm's estimated depth values seem a plausible basis for predicting human depth judgments, and the algorithm's fidelity score seems a plausible basis of rigidity judgments. An immediate problem is that perspective-dependent modulations in position of elements on the image plane are treated as noise by this (and most other) algorithms (Sperling & Dosher, 1987), although our subjects' depth percepts are improved by moderate amounts of perspective. The problems surrounding predictions with parallel and perspective projec-

tions are particularly enlightening. Detailed consideration (Dosher & Sperling, 1988; Sperling & Dosher, 1987) showed that a class of models including Ullman's (1984) exhibit unrealistic properties because they do not incorporate any perspective transformation, whereas the models would require a flexible perspective transformation to deal with perceptual facts. Parallel perspective algorithms, such as Ullman's, when applied to perspective images, such as those in our experiments, yield flattened depth estimates in relation to nonperspective images. On the contrary, for our human observers, perspective increased depth ratings slightly.

Predictions of perceived rigidity based on the fidelity criterion are the most problematic aspect of Ullman's (1984) algorithm. When the image is produced by perspective transformation, parallel-perspective rigidity algorithms (e.g., Ullman) cannot distinguish between veridical and reversed depth 3-D recovered objects—they yield precisely equal rigid and nonrigid solutions. For our subjects, reversed depth perceptions are grossly more nonrigid than the veridical ones, a powerful perceptual fact that is beyond the scope of this class of models.

Purely geometric algorithms that yield explicit solutions to 3-D objects given m views of n points (Bennett & Hoffman, 1985; Hoffman & Bennett, 1985; Ullman, 1979; Webb & Aggarwal, 1981) fare much worse than the incremental rigidity algorithm. Again, segmentation is simply assumed. The algorithms yield exact solutions under certain conditions in which the stimuli represent rigid objects. The outputs here are exact solutions or the fact that a solution failed. An exact solution must be rigid, so the model cannot predict any particular nonrigid percept, nor does it have computational by-products that can support rigidity–nonrigidity judgments, partial depth, or incomplete segmentation.

We conclude that, although many existing algorithms are of great interest as a possible basis for robotics solutions to the structure-from-motion problem, they are inadequate as psychological models. Our experiments suggest that a successful psychological model must identify at least three separable aspects of recovered objects that can serve as a basis for the three separable, measurable aspects of kinetic depth perception.

References

Andersen, G. J., & Braunstein, M. L. (1983) Dynamic occlusion in the perception of rotation in depth. *Perception & Psychophysics, 34,* 356–362.

Bennett, B. M., & Hoffman, D. D. (1985). The computation of structure from fixed-axis motion: Nonrigid structures. *Biological Cybernetics, 51,* 293–300.

Braunstein, M. L. (1962). Depth perception in rotating dot patterns: Effects of numerosity and perspective. *Journal of Experimental Psychology, 64,* 415–420.

Braunstein, M. L. (1977). Perceived direction of rotation of simulated three-dimensional patterns. *Perception & Psychophysics, 21,* 553–557.

Braunstein, M. L., & Andersen, G. J. (1981). Velocity gradients and relative depth perception. *Perception & Psychophysics, 29,* 145–155.

Braunstein, M. L., & Andersen, G. J. (1984). Shape and depth perception from parallel projections of three dimensional motion *Journal of Experimental Psychology. Human Perception and Performance, 10,* 749–760.

Caelli, T. M (1979). On the perception of some geometric properties of rotating three dimensional objects. *Biological Cybernetics, 33,* 29–37.

Caelli, T. M. (1980) Amplitude, frequency and phase determinants of perceived rotations and rigidity in the kinetic depth effect *Biological Cybernetics, 36,* 213–219.

Clocksin, W. F. (1980) Perception of surface slant and edge labels from optical flow: A computational approach *Perception, 9,* 253–269.

Cutting, J. E. (1987). Rigidity in cinema seen from the front row, side aisle. *Journal of Experimental Psychology. Human Perception and Performance, 13,* 323–334.

Dosher, B. A , Landy, M. S , & Sperling, G. (in press) Kinetic depth effect and optic flow. I. 3D Shape from Fourier motion *Vision Research*

Dosher, B. A., & Sperling, G. (1988). *Predicting rigid and nonrigid perceptions* Unpublished manuscript.

Dosher, B. A., Sperling, G , & Wurst, S. A. (1986). Tradeoffs between stereopsis and proximity luminance covariance *Vision Research, 26,* 973–990.

Gillam, B. (1975). New evidence for "closure" in perception. *Perception & Psychophysics, 17,* 521–524.

Gillam, B. (1976). Grouping of multiple ambiguous contours. Towards an understanding of surface perception. *Perception, 5,* 203–209.

Green, B. F., Jr. (1961). Figure coherence in the kinetic depth effect *Journal of Experimental Psychology, 62,* 272–282.

Gregory, R. L. (1970). *The intelligent eye* New York. McGraw-Hill

Hildreth, E. C., & Grzywacz, N. M. (1986) The incremental recovery of structure from motion: Position vs. velocity based formulations. *Proceedings of the Workshop on Motion: Representation and Analysis, IEEE Computer Society #696, Charleston, South Carolina, May 7–9, 1986* (pp. 137–144). Los Angeles: IEEE Computer Society Press.

Hoffman, D. D , & Bennett, B. M. (1985). Inferring the relative three-dimensional positions of two moving points. *Journal of the Optical Society of America A, 2,* 350–355.

Koenderink, J. J., & van Doorn, A. J. (1986). Depth and shape from differential perspective in the presence of bending deformations. *Journal of the Optical Society of America A, 3,* 242–249

Kropfl, W. J. (1975) *Variable raster and vector display processor* Unpublished Technical Memorandum TM-75-1223-3. Murray Hill, NJ: Bell Telephone Laboratories.

Landy, M. S. (1987). A parallel model of the kinetic depth effect using local computations *Journal of the Optical Society of America A, 4,* 864–877.

Landy, M. S , Dosher, B. A., & Sperling, G. (1986) Assessing kinetic depth in multi-dot displays. *Bulletin of the Psychonomic Society, 19,* 23.

Landy, M. S , Sperling, G , Dosher, B. A , & Perkins, M. E (1987) Structure from what kinds of motion? *Investigative Ophthalmology and Visual Science, ARVO Supplement, 28,* 233.

Lappin, J. S., Doner, J. F., & Kottas, B. L. (1980) Minimal conditions for the visual detection of structure and motion in three dimensions. *Science, 209,* 717–719.

Petersik, J. T. (1979) Three-dimensional object constancy: Coherence of a simulated rotating sphere in noise. *Perception & Psychophysics, 25,* 328–335.

Petersik, J. T. (1980). The effects of spatial and temporal factors on the perception of stroboscopic rotation simulations. *Perception, 9,* 271–283.

Schwartz, B., & Sperling, G. (1983). Luminance controls the perceived 3-D structure of dynamic 2-D displays. *Bulletin of the Psychonomic Society, 21*, 456–458.

Sperling, G., & Dosher, B. A. (1987). Predicting rigid and nonrigid perceptions. *Investigative Ophthalmology and Visual Science, ARVO Supplement, 28*, 362.

Sperling, G., Landy, M. S., Dosher, B. A., & Perkins, M. E. (1989). Kinetic depth effect and identification of shape. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 826–840.

Todd, J. T. (1984). The perception of three-dimensional structure from rigid and non-rigid motion. *Perception & Psychophysics, 36*, 97–103.

Ullman, S. (1979). *The interpretation of visual motion*. Cambridge, MA: MIT Press.

Ullman, S. (1984). Maximizing rigidity: The incremental recovery of 3-D structure from rigid and non-rigid motion. *Perception, 13*, 255–274.

Wallach, H., & O'Connell, D. N. (1953). The kinetic depth effect. *Journal of Experimental Psychology, 45*, 205–217.

Webb, J. A., & Aggarwal, J. K. (1981). Visually interpreting the motion of objects in space. *Computer, 14*, 40–49.

Received April 13, 1987
Revision received October 13, 1988
Accepted October 14, 1988 ■

Call for Nominations

The Publications and Communications Board has opened nominations for the editorships of the *Journal of Experimental Psychology: Animal Behavior Processes, Contemporary Psychology*, the Personality Processes and Individual Differences section of the *Journal of Personality and Social Psychology, Psychological Assessment: A Journal of Consulting and Clinical Psychology*, and *Psychology and Aging* for the years 1992–1997. Michael Domjan, Ellen Berscheid, Irwin Sarason, Alan Kazdin, and M. Powell Lawton, respectively, are the incumbent editors. Candidates must be members of APA and should be available to start receiving manuscripts in early 1991 to prepare for issues published in 1992. Please note that the P&C Board encourages more participation by members of underrepresented groups in the publication process and would particularly welcome such nominees. To nominate candidates, prepare a statement of one page or less in support of each candidate.

- For *JEP: Animal*, submit nominations to Bruce Overmier, Department of Psychology-Elliott Hall, University of Minnesota, 75 East River Road, Minneapolis, Minnesota 55455. Other members of the search committee are Donald A. Riley, Sara J. Shettleworth, Allan R. Wagner, and John L. Williams.
- For *Contemporary Psychology*, submit nominations to Don Foss, Department of Psychology, University of Texas, Austin, Texas 78712. Other members of the search committee are Edward E. Jones, Gardner Lindzey, Anne Pick, and Hans Strupp.
- For *JPSP: Personality*, submit nominations to Arthur Bodin, Mental Research Institute, 555 Middlefield Road, Palo Alto, California 94301. Other members of the search committee are Charles S. Carver, Ravenna S. Helson, Walter Mischel, Lawrence A. Pervin, and Jerry S. Wiggins.
- For *Psychological Assessment*, submit nominations to Richard Mayer, Department of Psychology, University of California-Santa Barbara, Santa Barbara, California 93106. Other members of the search committee are David H. Barlow and Ruth G. Matarazzo.
- For *Psychology and Aging*, submit nominations to Martha Storandt, Department of Psychology, Washington University, St. Louis, Missouri 63130. Other members of the search committee are David Arenberg and Ilene C. Siegler.

First review of nominations will begin January 15, 1990.

Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 2985–2989, April 1989
Psychology

Two motion perception mechanisms revealed through distance-driven reversal of apparent motion

CHARLES CHUBB AND GEORGE SPERLING

Human Information Processing Laboratory, Department of Psychology, New York University, 6 Washington Place, New York, NY 10003

Contributed by George Sperling, December 30, 1988

ABSTRACT **We demonstrate two kinds of visual stimuli that exhibit motion in one direction when viewed from near and in the opposite direction from afar. These striking reversals occur because each kind of stimulus is constructed to simultaneously activate two different mechanisms: a short-range mechanism that computes motion from space–time correspondences in stimulus luminance and a long-range mechanism in which motion computations are performed, instead, on stimulus contrast that has been full-wave rectified (e.g., on the absolute value of contrast).**

We demonstrate two dynamic visual stimuli that appear to move in one direction when viewed from near and in the opposite direction from afar. This remarkable reversal of apparent motion occurs because the stimuli are constructed to simultaneously activate two different mechanisms: a first-order mechanism that computes motion from space–time correspondences in raw stimulus *luminance* and a second-order mechanism that uses, instead, a *full-wave* rectified transformation (e.g., the absolute value) of stimulus *contrast* to compute motion.

The first stimulus, B, a rightward stepping, contrast-reversing bar, is a variant of Anstis's (1) reversed-phi stimulus. What we add are quite different explanations of the ordinary and the reversed motions in this stimulus and the conditions under which each is perceived.

The second stimulus, Γ, a stepping, contrast-reversing grating, is an elaboration of the first with two useful properties: (*i*) It provides the first- and second-order systems with motion signals of identical spatial frequency, moving at the same rate, but in opposite directions; and (*ii*) its motion direction is totally ambiguous to any half-wave rectifying system. The dominance of the first-order mechanism when the retinal image is small (far-viewing) suggests that it is the mechanism of Braddick's (2) short-range system; the dominance of the second-order mechanism with large retinal images suggests that it is the mechanism of the long-range system.

Since Braddick (2) proposed that there are two motion perception mechanisms with different properties—a short-range and long-range motion-perception system, the issue has been intensely investigated (3–16). The following differences between the short-range and long-range systems are proposed. The short-range system requires successive stimuli to be displaced in space by a small distance Δx within a small time period Δt and presented to the same eye. The long-range system tolerates large Δx, Δt, and interocular presentation (2, 12).

Anstis and Mather (16) noted that in making its matches across time and space, the long-range system is indifferent to sign of contrast. Motion is generated between successively displayed, spatiotemporally displaced points on a grey background, even when they are of opposite contrast polarity (i.e., one is white and the other black). Quite the reverse is true of the short-range system. The sensitivity of the short-range system to the sign of contrast is exhibited strikingly in the phenomenon of *reversed-phi apparent motion* (1): When a picture is flashed twice in quick succession, with the second flash slightly displaced in space from the first, motion (called ϕ motion) is perceived in the direction of the displacement. However, if the contrast of the picture is reversed between the first and second flash, motion may be perceived in the direction *opposite* to the displacement. This is *reversed-phi* motion.

What has been lacking is a clear specification of the mechanisms governing the short- and long-range systems. Here we introduce two stimuli, the contrast reversing bar B (Fig. 1*d*) and the stepping, contrast-reversing grating Γ (see Fig. 2*a*) that display short-range (reversed-phi) motion to the left when viewed from far away and long-range motion to the right when viewed from a short distance. Γ is constructed so as to place important constraints on the underlying mechanisms that detect the motion it displays from both far and near viewing distances. Specifically, Γ rules out the possibility that either sort of motion is mediated by *half-wave* rectification. Rather, Γ strongly suggests that the short-range system applies what we shall call standard motion analysis to raw stimulus *luminance*, while the particular long-range system stimulated by Γ from short viewing distances applies standard motion analysis to a *full-wave* rectified transformation of stimulus *contrast*.

A monochromatic visual stimulus is a function that assigns a luminous flux to each point in space–time. However, from a perceptual point of view, a stimulus is better described by its contrast than by its luminance *I*. Thus, a stimulus *S* is the normalized deviation of $I(x, y, t)$ from its mean luminance I_0, that is, for any point x, y, t in space–time, $S(x, y, t) = [I(x, y, t) - I_0]/I_0$. Because a stimulus is defined in terms of the contrast-modulation function *S* (rather than the raw luminance function *I*), stimulus values (unlike luminance values) may be positive or negative.

To simplify the discussion, we consider only stimuli that do not vary in the vertical dimension, i.e., stimuli that can be described as horizontally moving patterns of vertically oriented bars. Any such vertically-constant stimulus is characterized in all relevant respects by its xt cross-section $S(x, t)$, a slice made perpendicular to the vertical axis of space to reveal stimulus contrast as a function of horizontal space (x) and time (t).

Fig. 1*a* depicts eight frames of a movie of a dark vertical bar stepping left-to-right across a bright field. Fig. 1*b* is the xt cross-section of the rightward-stepping dark bar. Fig. 1*c* shows an xt cross-section of a rightward-drifting, vertically oriented sine-wave grating $S(x, t) = \sin(x - t)$. This sine-wave component of *b* is shown superimposed on *b*. Fig. 1*c* illustrates how the detection of motion in a complex stimulus can be understood in terms of motion of the sine-wave components.

It is immediately obvious from the xt cross-sections of the rightward-stepping bar and sine-wave stimuli that the prob-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Atch- 19

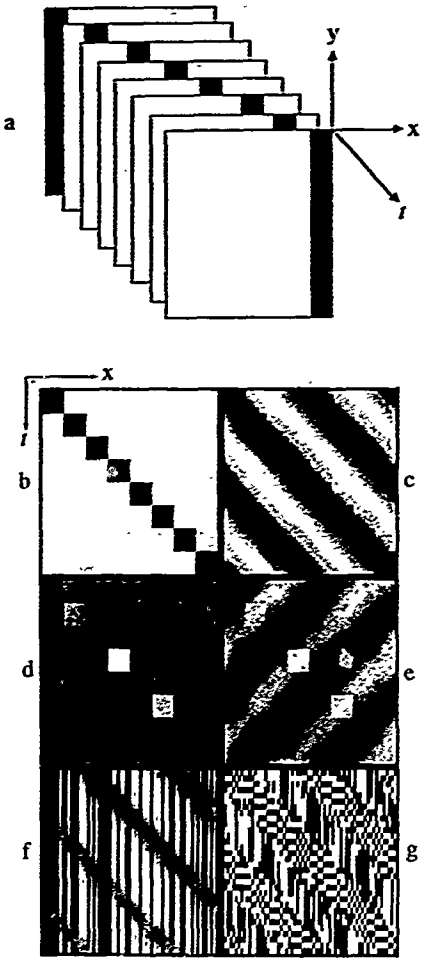


FIG. 1. Slant in x and y corresponds to motion in xt. (*a*) Eight frames in a display of a rightward-stepping, vertical bar: x and y represent the spatial dimensions of the display, and t represents time. This stimulus does not vary in the y dimension. Each of panels *b*–*g* is an xt cross-section of a dynamic stimulus that does not vary in y. (*b*) An xt cross-section of the rightward-stepping bar of panel *a*. Horizontal luminances are indicated along x, temporal luminances are indicated vertically with time t running downward. (*c*) Stimulus of *b* shown together with one of its largest sinusoidal components (*d*) The Gaussian windowed, contrast-reversing stepping bar—stimulus B. (*e*) Stimulus B shown together with its largest sinusoidal component, indicating why its far view (first-order, Fourier) motion is to the left (*f*) A row of vertical bars, randomly of positive or negative contrast the amplitude of which is modulated by a rightward-drifting grating (*g*) a row of random, black/white vertical bars the flicker rate of which is modulated by a rightward-drifting grating. The motion of all four stimuli is as obvious to all viewers as is the slant of the x, t cross-section The rightward motion of the black bar (*b* and *c*), and

lem of detecting motion in xt is equivalent to the problem of detecting orientation in xy. That is, the perception of an xy pattern slanting down to the right is analogous to the perception of an xt pattern moving to the right.

Fig. 1*d* shows the contrast-reversing bar B, which is Gaussian-windowed in time. When the bar takes 60 steps per sec (one step every 17 msec) and is windowed by a Gaussian function with SD of 25 msec, every observer so far has reported the direction of motion as being to the right when viewed in central vision from a wide range of near distances. On the other hand, B appears to be moving leftward when viewed in peripheral vision from near or when viewed in central vision from afar over a smaller range of distances near the vanishing point. Fig. 1*e* suggests the Fourier basis of the far-view motion; the dominant sine-wave components are leftward (17). We momentarily defer the explanation of rightward movement.

Visual slant detection (often called orientation detection) is generally thought to involve oriented Hubel–Wiesel (18) receptive fields in area 17 of the visual cortex. The corresponding computational mechanisms are oriented linear filters (19, 20). The detection of slant, however, involves a further (inherently nonlinear) stage of processing: A decision about the dominant slant of a spatial stimulus S must be made with reference to the relative energy in the responses-to-S of various linear filters in different phases and orientations. A wide range of models to explain slant (and motion) perception apply computations of this sort to the visual stimulus (17, 21–28), and similar computations are coming to have wide applications in robotic vision (29, 30).

Although exclusively spatial detectors are physically different from spatiotemporal detectors, the computations for orientation-detection and motion-detection are quite similar. For both slant and motion, the quantity computed by any energy-analytic detector can be cast as a linear combination of the pairwise products of stimulus values, $S(x_i, t_i)S(x_j, t_j)$, for i and j both ranging over all points in space-time. (For slant detectors the time variables t_i and t_j are replaced by vertical space variables y_i and y_j.) We refer to computations of this sort as *standard motion* (or *slant*) *analysis*.

Let D_1 be a standard motion analyzer defined for any stimulus S by

$$D_1(S) = \sum_i \sum_j W_{ij} S(x_i, t_i) S(x_j, t_j). \qquad [1]$$

where each W_{ij} is a real-valued weight. The standard motion analyzer tuned to the same sort of motion as D_1, but in the opposite direction, is

$$D_2(S) = \sum_i \sum_j W_{ij} S(x_i, t_j) S(x_j, t_i). \qquad [2]$$

Any stimulus S is called microbalanced if and only if for any such oppositely tuned standard motion analyzers, D_1 and D_2, the expected response $E[D_1(S)]$ is equal to the expected response $E[D_2(S)]$ (31).

Although, as this definition indicates, microbalanced random stimuli yield no signs of systematic motion to standard motion analysis, it is nonetheless possible to construct a wide variety of microbalanced random stimuli that display consistent motion across independent realizations (31, 32). For example, the amplitude-modulated noise stimulus I and the frequency-modulated noise stimulus J in Fig. 1 f and g (31) are microbalanced Nonetheless, observers universally perceive

the leftward far-view motion of the contrast-reversing bar B (*d* and *e*) are accessible to first-order mechanisms, the rightward motion of stimuli (f and g), and the rightward near-view motion of B (*d*) are not The motion of stimulus f can be exposed to standard analysis by simple half- or full-wave rectification, stimulus g requires a temporal linear filter (e g , a temporal differentiator) before rectification

the dynamic versions of these stimuli as moving rightward, and the texture versions as slanted downward to the right.

Following Cavanagh†, we call any motion mechanism that applies standard motion (or slant) analysis directly to luminance (or to a linear transformation of luminance) a first-order mechanism. Any motion mechanism that applies standard motion (or slant) analysis to a grossly nonlinear transformation of luminance is called a second-order mechanism.

There is a simple way to expose the inherent motion (or slant) in stimuli such as *I* and *J*: (*i*) apply a temporal (or vertical) linear filter, (*ii*) rectify the result, and (*iii*) apply standard analysis.‡ There are two candidate schemes of rectification: full-wave rectification, which consists of computing the absolute value (or a monotonically increasing function of the absolute value) of the filtered contrast, and half-wave rectification, which consists of making independent, separate computations on positive and on negative values of filtered contrast. Full-wave rectification has a long history of utility in signal processing. Half-wave rectification appears to be a widespread, almost universal, physiological process: Because neurons have only a positive output (their firing frequency), they are paired in order to economically convey positive and negative signal values. In the visual system, one pair-member (an "on-center" neuron) carries values of positive contrast, whereas its pair-mate (an "off-center" neuron) carries negative contrast values.

The phenomenon of reversed-phi motion (1) demonstrated in the far viewing of Fig. 1*d* [and many similar results (17)] could not occur if the short-range system applied a full-wave rectifier before standard motion analysis. Simple full-wave rectification of contrast obliterates the difference between the simple moving bar (Fig. 1*b*) and corresponding contrast-reversing bar (B, Fig. 1*d*). Any mechanism that full-wave rectified contrast before motion analysis would issue similar responses for the stimuli of Fig. 1 *b* and *d*.

These considerations do not, however, rule out the possibility that the short-range system uses a half-wave rectifier before standard motion analysis (33). Perhaps both short-range motion and the motion of various microbalanced random stimuli such as *I* (Fig. 1*f*) and *J* (Fig. 1*g*) can be explained with reference to a single kind of mechanism, one that applies to stimulus contrast a linear filter, then a half-wave rectifier, and finally some form of standard motion analysis. Or perhaps, as seems more likely, short-range motion results from applying standard motion analysis directly to contrast. In this case, we are left with the question of what sorts of rectification are involved in perceiving the motion of microbalanced random stimuli.

These issues are cleared up by the leftward-stepping, contrast-reversing grating Γ defined in Fig. 2. An *xt* cross-section of Γ is shown in Fig. 2*a*. The temporal scale and the (distance-dependent) spatial scale of the display are described in the legend for Fig. 2. Γ is perceived to move leftward from near viewing distances and rightward from far distances. Γ has been viewed by dozens of subjects in our lab, and the reversal of apparent motion with viewing distance has been observed by all.

The far-view motion of Γ is detected by the short-range system. Note that in each successive display, Γ is shifted 1/4 spatial cycle leftward, and its contrast is reversed. Thus, we should expect Γ to elicit reversed-phi motion under appropriate conditions. And indeed, when the spatial displacement between successive displays is made sufficiently small by moving the viewer back from the screen, Γ exhibits reversed-phi motion to the right, implicating the short-range system. The velocity of this far-view motion is easily distinguished by all subjects and is equal to that of the grating overlaid on Γ in Fig. 2*b*. As this overlay makes clear, the far-view motion of Γ is signaled directly by the distribution of energy in the Fourier transform of Γ. Typically, standard motion-analytic computations reflect this distribution of Fourier energy in the stimulus. Thus, the far-view motion of Γ is the predicted response of a first-order mechanism.

FIG. 2. Graphic analysis of the motion content of stimulus Γ, a horizontally windowed, leftward-stepping grating of vertical bars that reverses contrast with each step. (*a*) An *xt* cross-section of Γ. Γ is temporally periodic. The temporal slice displayed here contains eight frames, each of which lasts 1/15 sec; thus, the total duration shown is 533 msec. From far (8 m), the width of *a* is ≈0.6 degrees of visual angle (dva), and each vertical bar in the grating has a width of 0.02 dva. (*b*) A sinusoid is overlaid on Γ to illustrate the perceived motion of Γ when viewed from 8 m. Conformity to sinusoidal analysis suggests that the far-view motion of Γ is first-order. (*c*) |Γ|, the absolute value (full-wave rectified) transformation of Γ. From near (2 m), the stimulus Γ displays motion conforming to the sinusoid overlaid on |Γ|, suggesting that the near-view motion of Γ is second-order and possibly mediated by full-wave rectification of stimulus contrast.

By contrast, the near-view motion of Γ is detected by a second-order mechanism. It is evident to all viewers that the leftward motion displayed by Γ from short viewing distances is carried directly by the leftward-stepping, contrast-reversing, vertical bars. However, Γ has no energy in any Fourier component (drifting sinusoidal grating) whose velocity matches that of these leftward-stepping bars. This indicates that the near-view motion of Γ is not obtained directly by standard analysis. We can, however, expose the near-view motion of Γ by full-wave rectifying Γ before standard motion analysis. This is illustrated by Fig. 2*c*, in which |Γ| is shown, overlaid by a leftward-drifting grating that contributes strongly to it. The velocity of this sinusoid is precisely the velocity of the near-view motion of Γ.

There are other transformations aside from simple full-wave rectification that might expose the near-view motion of Γ to standard analysis. The most likely transformations (31, 32, 34) involve an initial stage of temporal linear filtering. Plausible candidates are filters whose response at every point (x,y) in space depends on (*i*) average recent stimulus contrast at that point and/or (*ii*) recent changes in contrast at that point. In particular, the likely temporal filters are marked by brief impulse responses (most of their energy confined to <100 msec) that (*i*) integrate to a nonzero value (so as to reflect raw stimulus contrast) and/or (*ii*) are biphasic (so as to register quick changes in contrast). Some candidate impulse responses are plotted in Fig. 3, *a*–*c*.

What distinguishes the leftward-stepping, contrast-reversing grating Γ from other stimuli that reverse direction of motion with viewing distance (34) is that, for all of these empirically plausible temporal linear filters $f*$ (e.g., with impulse response f conforming to Fig 3 *a*, *b*, or *c*), the result of half-wave rectifying $f*\Gamma$ is completely ambiguous in motion content.

Half-wave ambiguity of Γ and its transformations is illustrated in Fig. 3. The filter $g*$, whose impulse response g is shown in Fig 3*a*, is a physiologically plausible representation of the identity transformation, $g*$ averages recent contrast but does not register sudden changes in contrast. The filter $h*$, whose impulse response h is shown in Fig. 3*c*, is a

†Cavanagh, P., Conference on Visual Form and Motion Perception: Psychophysics, Computation, and Neural Networks, March 5, 1988, Boston University, Boston, MA.

‡Rectification alone suffices to expose the motion of *I* to standard analysis; temporal differentiation and rectification are required for *J*.

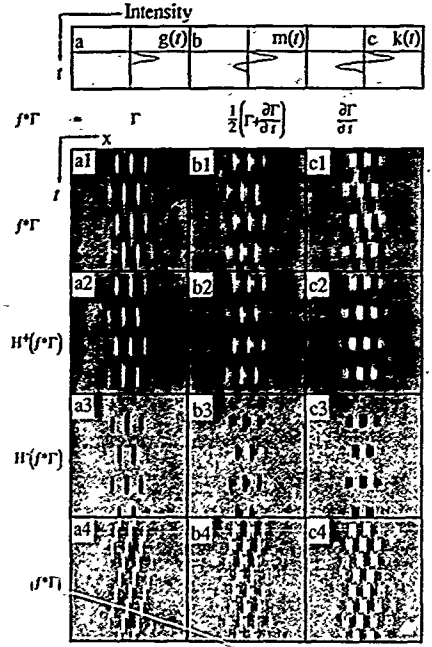


FIG 3 Exposing the near-view motion of Γ to standard analysis. The vertical dimension in all panels is time *t*, running downward. The scale of *t* is constant throughout the figure (see below) Each panel in the first row represents the impulse response *f* of a temporal filter that is an empirically plausible initial stage of a rectifying, second-order motion mechanism. The horizontal axis of each panel in the first row indicates intensity, increasing left-to-right. (*a*) Impulse response of a physiologically plausible approximation to the temporal identity. it averages recent stimulus contrast (*c*) A physiologically plausible approximation to a temporal differentiator. it responds only to temporal changes in contrast. (*b*) The average of responses of filters *a* and *c*, a physiologically plausible compromise between temporal differentiator and identity that indicates both recent changes in contrast and recent average contrast. The panels (*a1–c4*) are *xt* cross-sections, the horizontal axes indicate horizontal space, the vertical axes indicate time. Each grey panel spans 2 4° of visual angle horizontally at a viewing distance of 2 m and spans 533 msec (vertically). In the row $f * \Gamma$ and the column under each impulse response is a *xt* cross-section of the result of applying filter *f* to Γ (Fig. 2*a*) Subsequent rows indicate the result of rectifying $f * \Gamma$. $H^+(f * \Gamma)$ and $H^-(f * \Gamma)$ indicate the positive and negative half-wave components of the same-column linear transformation, and the row marked $|f * \Gamma|$ shows full-wave rectifications of these temporal filterings of Γ All half-wave components are ambiguous in motion content; all full-wave rectifications yield unambiguous leftward motion to standard analysis.

physiologically plausible approximation to a temporal differentiator ($k*$ registers temporal changes in contrast, without keeping track of average recent contrast). The best-of-both-worlds filter $m*$ has impulse response $m = (g + k)/2$ shown in Fig 3*b* The reason for including this best-of-both-worlds filter is that among the stimuli that display second-order motion mediated by temporal filtering, there are some for which $g*$ (Fig. 3*a*) works but not $k*$ (Fig. 3*c*), and some for which $k*$ (Fig. 3*c*) works but not $g*$ (Fig. 3*a*); however, $m*$ (Fig. 3*b*) works for all (32).

In Fig. 3, the top row of *xt* cross-sections (marked $f * \Gamma$) displays the result of applying each of the filters directly to Γ. The rows marked $H^+(f * \Gamma)$ (Fig. 3 *a2*, *b2*, and *c2*) and $H^-(f * \Gamma)$ (Fig. 3 *a3*, *b3*, and *c3*) display the positive and negative half-wave components of the same-column, filtered outputs (Fig. 3 *a1*, *b1*, and *c1*), and the row marked $|f * \Gamma|$ displays full-wave rectifications (Fig. 3 *a4*, *b4*, and *c4*) of the filter outputs. The important fact graphically illustrated here is that the half-wave components of all of these linear transformations of Γ are completely ambiguous in motion content. As Fig. 3 *a4*, *b4*, and *c4* make clear, full-wave rectification works to expose the near-view motion of Γ; however, almost any full-wave-like rectification that combines same-sign output for positive and negative signal components will also work.

The distance-driven reversal of the apparent motion displayed by the leftward-stepping, contrast-reversing grating Γ (Fig. 2*a*) makes it dramatically clear that, as many have observed (2, 16, 30–38), the visual system extracts motion information from the visual signal in more than one way. Fig. 2 *b* and *c* illustrate that the far-view motion of Γ is consonant with a first-order mechanism (i.e., a Fourier mechanism that applies some form of standard motion analysis directly to the untransformed stimulus), whereas the near-view motion of Γ implicates a second-order mechanism that applies standard motion analysis to a rectified transformation of Γ (e.g., |Γ|, Fig. 2*c*). In the context of the various stimuli we have been able to create, the motion of which reverses with distance (33), the specific importance of Γ derives from the fact that the near-view motion of Γ cannot be exposed to standard motion analysis by any of the empirically plausible linear filters followed by half-wave rectification, whereas full-wave rectification works in conjunction with all the plausible filters.

It is possible to construct stimuli the motion of which is accessible neither to first-order mechanisms nor to any of the second-order mechanisms considered here (32). The question remains open as to whether any of the mechanisms that detect these other sorts of motion use half-wave rectification. However, the leftward-stepping, contrast-reversing grating Γ conclusively establishes that at least one second-order mechanism uses full-wave rectification.

This work was supported by Air Force Office of Scientific Research, Life Sciences Directorate, Vision Information Processing Program, Grants 85-0364 and 88-0140.

1 Anstis, S. M. (1970) *Vision Res* 10, 1411–1430.
2 Braddick, O. (1974) *Vision Res.* 14, 519–527.
3 Lappin, J S & Bell, H. H. (1976) *Vision Res* 16, 161–168.
4 Westheimer, G. & McKee, S. P. (1977) *Vision Res.* 17, 887–892.
5 Bell, H H & Lappin, J S. (1979) *Percept. Psychophys.* 26, 415–417.
6 Baker, C L & Braddick, O (1982) *Vision Res* 22, 851–856.
7 Baker, C L. & Braddick, O (1982) *Vision Res* 22, 1253–1260
8. Chang, J J. & Julesz, B. (1983) *Vision Res.* 23, 639–646
9 Chang, J J & Julesz, B. (198-) *Vision Res* 23, 1379–1386.
10. Chang, J J & Julesz, B. (1983) *Spatial Vision* 1, 39–45
11 Ramachandran, V S. & Anstis, S. M. (1983) *Vision Res.* 23, 1719–1724.
12. Westheimer, G. (1983) *Vision Res.* 23, 759–763.
13 Bennett, R. G & Westheimer, G (1985) *Vision Res* 25, 565–569
14 Nakayama, K. & Silverman, G. (1984) *Vision Res* 24, 293–300
15 van Doorn, A. J. & Koendennk, J. J. (1984) *Vision Res.* 24, 47–54.
16 Anstis, S M & Mather, G. (1985) *Perception* 14, 167–179.
17 van Santen, J P H & Sperling, G. (1984) *Opt. Soc. Am. A* 1, 451–473.
18 Hubel, D H & Wiesel, T N (1959) *J Physiol. (London)* 148, 574–591.
19 Granlund, G H & Knutsson, H. (1983) in *Physical and Biological Processing of Images*, eds. Braddick, O. J. & Sleigh, A. C. (Springer, New York), pp. 282–303.

20 Watson, A. B. (1983) in *Physical and Biological Processing of Images*, eds. Braddick, O. J. & Sleigh, A C. (Springer, New York), pp. 100–114.
21. Reichardt, W. (1957) *Z. Naturforschung* 12, 447–457.
22. Adelson, E. H. & Bergen, J. (1985) *J. Opt. Soc. Am. A* 2, 284–299.
23 Watson, A. B. & Ahumada, A. J. (1983) *A Look at Motion in the Frequency Domain* NASA Technical Memorandum 84352 (Natl. Aeronautics and Space Admin . Ames Research Center, CA).
24. Watson, A. B. & Ahumada, A. J. (1985) *J. Opt. Soc. Am A* 2, 322–342.
25. van Santen, J P. H & Sperling, G. (1985) *J. Opt. Soc. Am A* 2, 300–321.
26. Fleet, D. J. & Jepson, A. D. (1985) *On the Hierarchical Construction of Orientation and Velocity Selective Filters*, Technical Report RBCV-TR-85-8 (Univ. Toronto Comp. Sci. Dept., Toronto. ON).
27. Heeger, D. J. (1987) *J. Opt. Soc. Am A* 4, 1455–1471.
28 Reichardt, W. & Egelhaaf, M. (1988) *Z. Naturwissenschaften* 75, 313–315.
29. Anandan, P. (1988) in *Proceedings of the First International Conference on Computer Vision* (IEEE Comp. Soc., Washington, DC), pp. 219–230.
30. Waxman, A. M. & Bergholm, F. (1987) *Convected Activation Profiles and Image Flow Extraction*, Laboratory for Sensory Robotics Technical Report 4, (Boston Univ., Boston, MA).
31. Chubb, C. & Sperling, G. (1988) *J. Opt. Soc. Am. A* 5, 1986–2007.
32. Chubb, C. & Sperling, G. (1989) *Proceedings: 1989 IEEE Workshop on Motion*, (IEEE Computer Society, Washington, DC).
33. Walt, R. J. & Morgan, M J. (1985) *Vision Res.* 25, 1661–1674
34. Chubb, C. & Sperling, G. (1988) *Invest. Ophthalmol. Vis. Sci.* 29, 266.
35. Chubb, C. & Sperling, G. (1988) *Mathematical Studies in Perception and Cognition*, 88-1 (New York Univ., New York).
36. Chubb, C. & Sperling, G. (1987) *Invest. Ophthalmol. Vis. Sci.* 28, 233.
37. Pantle, A & Picciano, L. (1976) *Science* 193, 500–502.
38. Lelkens, A. M. M. & Koenderink, J. J. (1984) *Vision Res.* 24, 1083–1090.

George Sperling and Charles Chubb. Apparent Motion Derived From Spatial Texture. Investigative Opthalmology and Visual Science, 1989, **30, No. 3,** *ARVO Supplement, 425*

APPARENT MOTION DERIVED FROM SPATIAL TEXTURE

George Sperling and *Charles Chubb*, Human Information Processing Laboratory, New York University, New York, NY 10003.

Texture quilts are dynamic stimuli designed for studying motion-from-spatial-texture without contamination by motion mechanisms sensitive to other aspects of the signal. Here we provide a theoretical foundation and concrete stimulus-construction methods. We demonstrate texture quilts that exhibit strong apparent movement but whose motion content is unavailable to standard motion analysis such as might be accomplished by an Adelson/Bergen motion-energy analyzer, a Watson/Ahumada motion sensor, or by any Reichardt detector. Furthermore, the following transformations leave the motion in texture quilts unavailable to standard motion analysis: (a) any linear space-time separable transformation or (b) any purely temporal transformation, no matter how nonlinear (e.g., rectifying a temporal derivative). Applying (a) or (b) to a texture quilt results in a spatiotemporal function P (not necessarily a texture quilt) that is again microbalanced–that is, its motion is unavailable to standard motion analysis. The simplest mechanism sufficient to sense the motion exhibited by texture quilts consists of three successive stages: (i) a purely spatial linear filter (the "texture grabber"), (ii) a rectifier to transform regions of high-energy filter response into regions of high average value, and (iii) standard motion analysis. Stimuli of a still higher order that require a re-iteration of stages (i) and (ii) to yield motion will be demonstrated.

Supported by AFOSR Life Sciences, Visual Information Processing Program, Grant 88-0140.

Atch-20

Charles Chubb, George Sperling, and Joshua A. Solomon. Texture Interactions Determine Apparent Lightness. Investigative Opthalmology and Visual Science, 1989, 30, No. 8, *ARVO Supplement*, Pp. 1683

Erratum

The wrong abstract was printed on page 161, no. 11 of the March, 1989 Vol. 30, No. 3 Supplement to *Investigative Ophthalmology and Visual Science.* The correct abstract was actually presented and is printed below.

TEXTURE INTERACTIONS DETERMINE APPARENT LIGHTNESS

Charles Chubb, George Sperling, and *Joshua A. Solomon.*
Human Information Processing Laboratory, New York University.

We demonstrate that for a test patch of binary spatial noise P embedded in a surrounding noise field S, the perceived contrast of P depends substantially on the contrast of the noise surround S. When P is surrounded by high-contrast noise, its bright points appear dimmer, and simultaneously, its dark points appear less dark than when P is surrounded by a uniform field, even though local mean luminance is kept constant across all displays. Sinusoidally modulating the contrast C_S of the noise surround S causes the apparent contrast of P to modulate in antiphase to C_S. In a nulling experiment, C_S was modulated between 0 and 1 at 0.47 Hz. For noise patches P of mean contrast C_P between *0.3* and *0.5*, the amplitude of the induced modulation of P's apparent contrast was on the order of $0.45C_P$. By comparison, when the noises in P and S, respectively, are filtered into nonoverlapping, octave-wide spatial frequency bands, the modulation of C_S has very little effect on the apparent contrast of P. These results suggest that the perceived lightness or darkness of a point in space depends on the combined responses of multiple units at that point, where each unit is tuned to a specific band of spatial frequencies, and the response of each unit is normalized relative to the responses of nearby units of the same type.

Supported by AFOSR Life Sciences, Visual Information Processing Program, Grant 88-0140

Atch-21

Proceedings: Workshop on Visual Motion: 91 0757
(March 20-22, Irvine, California.)
Washington, D.C: IEEE COmputer Society Press, 1989. Pp. 126-138.

Second-Order Motion Perception: Space/time Separable Mechanisms

Charles Chubb George Sperling

Human Information Processing Laboratory, Department of Psychology,
New York University, 6 Washington Place, New York, NY 10003

Abstract

Microbalanced stimuli are dynamic displays which do not stimulate motion mechanisms that apply *standard* (Fourier-energy or autocorrelational) *motion analysis* directly to the visual signal. Because they bypass such *first-order* mechanisms, microbalanced stimuli are uniquely useful for studying *second-order* motion perception (motion perception served by mechanisms that require a grossly nonlinear stimulus transformation prior to standard motion analysis). Some stimuli are *microbalanced under all pointwise stimulus transformations* and therefore are immune to early visual nonlinearities. We use them to disable motion information derived from spatial (temporal) filtering in order to isolate the temporal (spatial) properties of space/time separable second-order motion mechanisms. The motion of all of the microbalanced stimuli we consider can be extracted by (1a) band-selective spatial filtering and (1b) biphasic temporal filtering, nonzero in dc, followed by (2) a rectifying nonlinearity and (3) standard motion analysis.

1. Introduction.

Standard motion analysis. A visual display is described by $L(x,y,t)$, its luminance as a function of space, x,y, and time, t. We use the term *standard motion analysis* for any computation applied to L *that derives L's motion from correlations of L-values across time and space.* Such computations are consonant with the *motion-from-Fourier-components principle*, which states that L's motion is reflected in some reasonable way by the contributions to L of individual Fourier components (drifting sinusoidal gratings). The recently proposed motion-perception theories of Adelson & Bergen [1], Heeger [5], van Santen & Sperling [3,4], and Watson & Ahumada [2] all perform various forms of standard motion analysis on their input. Similarly, the computer vision models of Anandan [9] and Waxman & Bergholm [10] also perform standard motion analysis on the input signal.

First-order mechanisms. A fundamental transformation generally presumed to be subjected to standard motion analysis in human visual processing is the *contrast* of the signal (the normalized deviation of luminance from its locally computed mean). We call mechanisms *first-order* that apply standard motion-analysis to raw stimulus contrast. Any motion mechanism that applies a grossly nonlinear transformation to the stimulus prior to standard motion analysis, we call *second-order*.

It is becoming clear, from apparently moving stimuli which do not stimulate standard motion detectors, that first-order mechanisms cannot account for all the data [11-28]. In particular, Chubb and Sperling [24,26,27] have demonstrated a variety of stimuli which display consistent, unambiguous apparent motion, yet which do not systematically stimulate first-order mechanisms.

The methods used by Chubb & Sperling [26] to construct apparent motion stimuli devoid of systematic first-order motion content are founded on the notion of a *microbalanced* random stimulus. A random stimulus I is *microbalanced* iff, for any space/time separable function W, the result $J=WI$ of multiplying I by W satisfies the following condition. (J is *drift balanced*) the expected power in J of any given drifting sinusoidal grating is equal to the expected power in J of the grating of the same spatial frequency, drifting at the same rate, but in the opposite direction. Drift-balanced and microbalanced random stimuli are useful for studying motion perception because they provide flexible access to *second-order* motion mechanisms without systematically engaging first-order mechanisms.

In this paper, we begin by reviewing the basic results about drift-balanced and microbalanced random stimuli, then apply these findings to generate a collection of microbalanced stimuli displaying various types of motion. The motion of each of the stimuli we consider is best revealed to standard analysis by a *space/time separable* linear filter followed by a rectifier. The first two microbalanced stimuli we discuss (stimuli 3.1 and 3.2) place important constraints on the *temporal* filtering mediating space/time separable, second-order motion-perception. The motion of each of the last four stimuli (4.2.2, 4.2.3, 4.2.5, and 4.2.6) depends only on the *spatial* filtering stage (temporal filtering alone, followed by rectification, cannot expose the motion of these stimuli).

A transformation is *pointwise* if its output value at a point (x,y,t) in space/time depends only on the value of the input at (x,y,t). Pointwise transformations include what are often called "static nonlinearities." Stimuli 3.2, 4.2.2, 4.2.3, 4.2.5 and 4.2.6 all *remain microbalanced after arbitrary pointwise transformations.* We present general methods for constructing stimuli of this sort.

Atch-22

A transformation is *purely temporal* if its output value at a point (x,y,t) depends only on the history of input at (x,y). The class of purely temporal transformations is very general and includes, for example, temporal bandpass filtering preceded and followed by arbitrary pointwise transformations. Stimuli 4.2.2, 4.2.3, 4.2.5 and 4.2.6 *remain microbalanced after any purely temporal transformation*. Such stimuli are extremely useful for investigating second-order motion perception, because they provide a critical measure of control in differentially stimulating specific second-order mechanisms. Indeed, under virtually all models of visual processing, the first effective transformation mediating the perception of motion displayed by such stimuli is bound to be a *spatial linear filter* (a "texture-grabber"). This linear stage must, of course, be followed by a pointwise nonlinearity (such as rectification or thresholding) to expose the microbalanced stimulus motion to standard analysis.

2. Preliminaries.

Section outline. In this section we state the background facts presupposed by the main discussion of the paper. The broad topics covered are:

- Real-valued, discrete visual stimuli and their Fourier transforms. We take a stimulus to be a real-valued function whose action is restricted to a finite grid of spatiotemporal sampling locations.

- Transformations. Definitions are given of linear shift-invariant transformations, and pointwise transformations.

- Random stimuli. A random stimulus is a jointly distributed set of random variables assigned to a grid of spatiotemporal sampling locations.

- Drift-balanced and microbalanced random stimuli. A random stimulus I is drift balanced iff the expected power contributed to I by any given Fourier component (drifting-sinusoidal grating) is equal to the expected power in I of the grating of the same spatial frequency drifting at the same rate, but in the opposite direction. I is microbalanced iff WI is drift balanced for any space/time separable function W that "windows" I. The class of microbalanced random stimuli is significant for studying motion-perception, since (i) it is easy to construct a broad range of microbalanced random stimuli which display consistent, compelling apparent motion across independent realizations, despite the fact that (ii) the motion displayed by any microbalanced random stimulus is invisible to first-order mechanisms, regardless of the spatiotemporal scope over which they perform their motion-analysis.

of f with g by $f * g$, and the product of f with g by fg.

2.1. Discrete dynamic visual stimuli and their Fourier transforms.

We let $\mathbf{R}$ denote the real numbers, and $\mathbf{Z}$ ($\mathbf{Z}^+$) the integers (positive integers).

Contrast modulation. Luminance $l(x,y,t)$ is physically constrainted to be a non-negative quantity. Psychophysically, the significant quantity is *contrast*, the normalized deviation at each time t of luminance at each point (x,y) in the visual field from l_o a "background level", or "level of adaptation", which reflects the average luminance over points proximal to (x,y,t) in space and time. We shall restrict our attention throughout this paper to stimuli for which it can be assumed that the background luminance level l_o is uniform over the significant spatiotemporal locations in the display.

For any stimulus l with base luminance l_o, call the function I satisfying

$$l = l_o(1 + I),$$

the *contrast modulator* of l (and note that $I \geq -1$).

Psychophysically, it is well-established that over substantial ranges of l_o, the apparent motion of I does not depend upon l_o. Therefore, we shift our focus from luminance to contrast, and identify a stimulus with its contrast modulator, dropping reference to background level.

Stimuli. We restrict ourselves to discrete stimuli, whose activity is restricted to a finite grid of points in space/time. Specifically, we call any function $I:\mathbf{Z}^3 \to \mathbf{R}$ a *stimulus* iff $I[x,y,t] = 0$ for all but finitely many points of $\mathbf{Z}^3$. We shall be considering stimuli as functions of two spatial dimensions and time. The reader may find it convenient to think of the first spatial dimension (always indexed by x) as horizontal, with values increasing to the right, the second spatial dimension (indexed by y) as vertical, with values increasing upward. The temporal dimension is indexed by t. For concreteness, the reader is encouraged to imagine $\mathbf{Z}^3$ as indexing the pixels in a dynamic digital display.

Because any stimulus I is nonzero at only a finite number of points, the power in I is finite, from which we observe that I has a well-defined Fourier transform.

We denote I's Fourier transform by $\bar{I}$:

$$\bar{I}(\omega,\theta,\tau) = \sum_{x,y,t \in \mathbf{Z}} I[x,y,t]e^{-j(\omega x + \theta y + \tau t)}.$$

Although $\bar{I}$ is defined for all real numbers ω,θ,τ, it is periodic over 2π in each variable. This fact is reflected in the inverse transform:

$$I[x,y,t] = \frac{1}{(2\pi)^3}\int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi} \bar{I}(\omega,\theta,\tau)e^{j(\omega x + \theta y + \tau t)}d\omega\, d\theta\, d\tau.$$

In the Fourier domain, ω indexes frequencies relative to x, θ indexes frequencies relative to y, and τ indexes frequencies relative to t.

We distinguish the stimulus $\mathbf{0}$ by setting $\mathbf{0}[x,y,t] = 0$ for all $x,y,t \in \mathbf{Z}$.

Any stimulus I is called *space/time separable* if $I[x,y,t] = g[x,y]\,h[t]$, for some real-valued functions g and h of space and time respectively.

2.2. Transformations.

Any function T which takes the set of real-valued functions of Z^3 into itself is called a *transformation*. If, for instance, $I: Z^3 \to R$ then $T(I): Z^3 \to R$, and we write $T(I)[x,y,t]$ to indicate the value of $T(I)$ at any point $(x,y,t) \in Z^3$. We shall be particularly concerned with two types of transformations: *linear shift-invariant* transformations, and *pointwise* transformations.

Pointwise transformations, rectifiers. For any functions $f: A \to B$ and $g: B \to C$, the *composition* $g \bullet f: A \to C$ is given by

$$g \bullet f(a) = g(f(a))$$

for any $a \in A$. Then for any $f: R \to R$, we call the transformation $f \bullet$, yielding the spatiotemporal function $f \bullet I$ when applied to stimulus I, a *pointwise* transformation (because its output value at any point in space/time depends only on its input value at that point). The transformation $f \bullet$ is called a *positive half-wave rectifier* if f is monotonically increasing, and $f[v] = 0$ for all $v \leq 0$. $f \bullet$ is called a *negative half-wave rectifier* if f is monotonically decreasing, and $f[v] = 0$ for $v \geq 0$. Finally, $f \bullet$ is called a *full-wave* rectifier if f is a monotonically increasing function of absolute value.

Linear, shift-invariant transformations. Linear, shift-invariant (LSI) transformations are spatiotemporal convolutions: For $k: Z^3 \to R$ (the *impulse response*), the LSI transformation $k*$ yields the convolution $k * I$ when applied to any stimulus I: i.e., for any $\alpha \in Z^3$,

$$k * I[\alpha] = \sum_{\beta \in Z^3} I[\beta]\, k[\alpha - \beta].$$

2.3. Random stimuli.

The notion of a random stimulus generalizes that of a nonrandom stimulus in that the values assigned points in space/time by a random stimulus are random variables rather than constants. A random stimulus is a family $\{R[x,y,t] \mid x,y,t \in Z\}$ of random variables, all but some finite number of which are always 0. To ensure that R has a well-defined expected power spectrum we require that $R[x,y,t]$ has a finite second moment for each $(x,y,t) \in Z^3$:

2.3.1. Call any family $\{R[x,y,t] \mid (x,y,t) \in Z^3\}$ of jointly distributed random variables a *random stimulus* provided

(i) for all but finitely many $(x,y,t) \in Z^3$, $R[x,y,t]$ is invariably equal to 0,

and

(ii) $E\left[R[x,y,t]^2\right]$ exists for all $(x,y,t) \in Z^3$.

As with non-random stimuli, we write $\bar{R}$ for the Fourier transform of the random stimulus R. R is called *space/time* separable iff R is space/time separable with probability 1. If there exists a stimulus S such that $R = S$ with probability 1, then R is called *constant*.

2.4. Drift-balanced and microbalanced random stimuli.

The *motion-from-Fourier-components principle* is a commonly encountered rule of thumb for predicting the apparent motion of an arbitrary stimulus $I[x,y,t] = f[x,t]$ that is constant in the vertical dimension of space. It states that, for I considered as a linear combination of drifting sinusoidal gratings, if the power in I of the rightward-drifting gratings is greater than the power of the leftward-drifting gratings, then apparent motion should be to the right. Conversely, if most of I's power resides in the leftward-drifting gratings, apparent motion should be to the left. Otherwise I should manifest no decisive motion in either direction.

This prediction rule for horizontally moving stimuli is a restricted version of the more general *motion-from-Fourier-components* principle: For any stimulus L to exhibit motion in a certain direction in the neighborhood of some point $(x,y,t) \in Z^3$, there must be some spatiotemporal volume Δ proximal to (x,y,t) such that the Fourier transform of L computed locally across Δ has substantial power over some regions of the frequency domain whose points correspond, in the space/time domain, to sinusoidal gratings drifting in a direction consistent with the motion perceived.

The following class of random stimuli provides a rich pool of counterexamples to the motion-from-Fourier-components principle [26].

2.4.1. Call any random stimulus R *drift balanced* iff

$$E\left[|\bar{R}(\omega,\theta,\tau)|^2\right] = E\left[|\bar{R}(\omega,\theta,-\tau)|^2\right]$$

for all $(\omega,\theta,\tau) \in R^3$.

Thus, a random stimulus R is drift balanced iff the expected power in R of each drifting sinusoidal component is equal to the expected power of the component of the same spatial frequency, drifting at the same rate, but in the opposite direction. That is, that expected power of every frequency is the same, independently of whether a series of frames is displayed in forward or reverse order. Obviously, for any class of spatiotemporal receptors tuned to stimulus power in a certain spatiotemporal frequency band, a drift-balanced random stimulus will, on the average, stimulate equally well those receptors tuned to the corresponding band, of opposite temporal orientation.

Microbalanced Random Stimuli. Consider the following two-flash stimulus S: In flash 1, a bright spot (call it Spot 1), appears. In flash 2, Spot 1 disappears, and two new spots appear, one to the left and one symmetrically to the right of Spot 1. As one might suppose, S is drift balanced. On the other hand, it is equally clear that a first-order motion detector whose spatial reach encompassed the location of Spot 1 and only *one* of the Spots in flash 2 might well be stimulated in a fixed direction by S. Thus, although S is drift balanced, some first-order motion detectors may be stimulated strongly and systematically by S. These detectors can be differentially selected by *spatial windowing*, and thereby the drift-balanced stimulus S can be converted into a non-drift-balanced stimulus by multiplying it by an appropriate space/time separable

function. This property is escaped by the following subclass of drift-balanced random stimuli.

2.4.2. Call any random stimulus I *microbalanced* iff WI is drift balanced for any space/time separable (non-random) function W.

One can think of the multiplying function W as a "window" through which a spatiotemporal subregion of I can be "viewed" in isolation. The space/time separability of W insures that it is "transparent" with respect to the motion content of the region of to which it is applied: W does not distort I's motion with any motion content of its own. Thus, the fact that I is microbalanced means that any subregion of I encountered through a "motion-transparent window" is drift balanced.

The following characterization of the class of microbalanced random stimuli, and the rest of the results in this section are from Chubb and Sperling [26].

2.4.3. *A random stimulus I is microbalanced if and only if*

$$E\Big[I[x,y,t]\,I[x',y',t'] - I[x,y,t']\,I[x',y',t]\Big] = 0$$

for all $(x,y,t),(x',y',t') \in Z^3$.

Some other relevant facts about microbalanced random stimuli:

2.4.4. *For any independent microbalanced random stimuli I and J,*

I. *the product IJ is microbalanced,*

and

II. *the convolution $I * J$ is microbalanced.*

2.4.5. *(a) Any spacetime separable random stimulus is microbalanced; (b) any constant microbalanced random stimulus is spacetime separable.*

The following result is useful in constructing a wide range of microbalanced random stimuli which display striking apparent motion.

2.4.6. *Let Γ be a family of pairwise independent, microbalanced random stimuli, all but at most one of which have expectation 0. Then any linear combination of Γ is microbalanced.*

The Reichardt detector characterization of microbalanced random stimuli. Two first-order motion detectors proposed for psychophysical data [1,6] can be recast as variants of a *Reichardt Detector* [3,4,31]. The Reichardt detector has many useful properties as a motion detector without regard to its specific instantiation [3,4].

Figure 1 shows a diagram of the Reichardt detector. The Reichardt detector consists of a left and a right subunit that share their inputs. The left subunit normally computes leftward motion because the filter g_1* acts as an internal delay to match the external delay of a moving stimulus. The right subunit normally computes rightward motion. The output represents the smoothed leftward minus rightward difference.

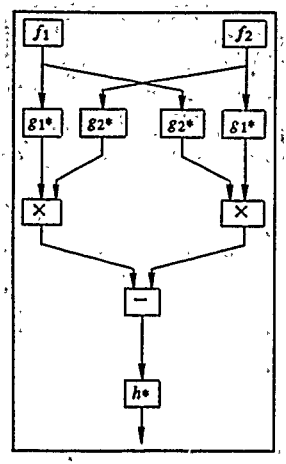


Figure 1: *The Reichardt detector.* The detector consists of a left and a right subunit; the left unit normally detects leftward movement; the right unit, rightward movement. In response to a stimulus I, each spatial input filter (receptive field) f_i outputs a temporal function that is then convolved with a temporal filter g_i*. The correlator boxes, marked "×", ouput the product of their inputs. The box marked "−" outputs its left input minus its right; this output indicates the net leftward minus rightward motion. The box $h*$ contains a temporal smoothing filter to produce time-averaged output.

Specifically, the Reichardt detector consists of spatial receptors characterized by spatial window functions (receptive fields) f_1 and f_2, temporal filters g_1* and g_2*, multipliers, a differencer, and another temporal filter $h*$. The spatial receptors f_i, $i = 1, 2$, act on the input stimulus I to produce intermediate outputs,

$$y_i[t] = \sum_{(x,y)\in Z^2} f_i[x,y]\,I[x,y,t].$$

At the next stage, each temporal filter g_j* transforms its input y_i $(i, j = 1, 2)$, yielding four temporal output functions: $g_j * y_i$. The left and right multipliers then compute the products

$$\Big[y_1 * g_1[t]\Big]\Big[y_2 * g_2[t]\Big] \text{ and } \Big[y_1 * g_2[t]\Big]\Big[y_2 * g_1[t]\Big]$$

respectively, and the differencer subtracts the output from the right multiplier from that of the left multiplier:

$$D[t] = \Big[y_1 * g_1[t]\Big]\Big[y_2 * g_2[t]\Big] - \Big[y_1 * g_2[t]\Big]\Big[y_2 * g_1[t]\Big]$$

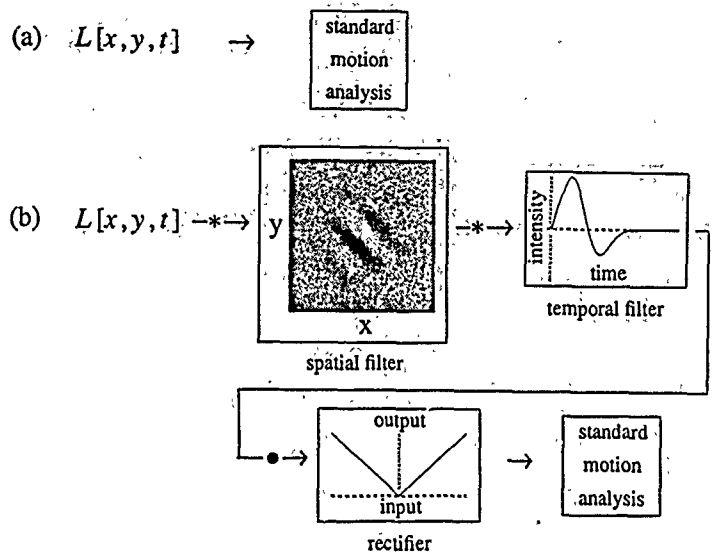


Figure 2: *First-order and second-order motion mechanisms.* (a) First-order motion mechanisms apply standard motion analysis (e.g., Reichardt model) directly to the luminance signal L. Many second-order mechanisms can be modeled by a signal transformation comprised of a spatiotemporal linear filter, followed by a pointwise nonlinearity followed by standard motion analysis. The filtering performed in (b) is *space/time separable* (spatial filtering and temporal filtering occur in separate boxes), followed by a pointwise nonlinearity, which is illustrated here with a full-wave rectifier. The motion of all the microbalanced stimuli considered in this paper can be extracted by the second-order mechanisms diagrammed in (b) with appropriately chosen spatial and temporal filters.

The final output is produced by applying the filter $h*$, whose purpose is to appropriately smooth the time-varying, differencer output D. Since almost all first-order mechanisms can be expressed as, or closely approximated by Reichardt detectors, the following result [27] is the cornerstone of the claim that microbalanced random stimuli bypass first-order motion mechanisms.

2.4.7. *For any random stimulus I, the following conditions are equivalent:*

(a) I is microbalanced.

(b) The expected response of any Reichardt detector to I is 0 at every instant in time.

Varieties of microbalanced motion.

In Sections 3 and 4, we describe six random stimuli, all of which are microbalanced, yet display consistent apparent motion across independent realizations. For each of these random stimuli I, the motion displayed by I can be exposed to standard motion analysis by a transformation

$$T(I) = r \bullet (f * I) , \qquad (1)$$

where $r \bullet$ is a rectifier, and $f*$ is a space/time separable filter.

3. Motion mediated by simple rectification and by temporal differentiation followed by rectification.

The first two stimuli (3.1 and 3.2) place constraints only on the temporal component of the the filter $f*$. Subsequent stimuli focus on the spatial component.

3.1. Stimulus: The amplitude-modulating squarewave. The motion of some of the microbalanced stimuli demonstrated by Chubb & Sperling [24,26] results from modulating the *amplitude* of spatially independent, visual noise. For example, Fig. 3a shows an xt cross-section of a squarewave, stepping 1/4 spatial-cycle leftward each frame, modulating (between 0 and 1) the amplitude of a row of static, horizontally independent black/white vertical bars. This stimulus displays obvious leftward motion to all viewers under a broad range of viewing

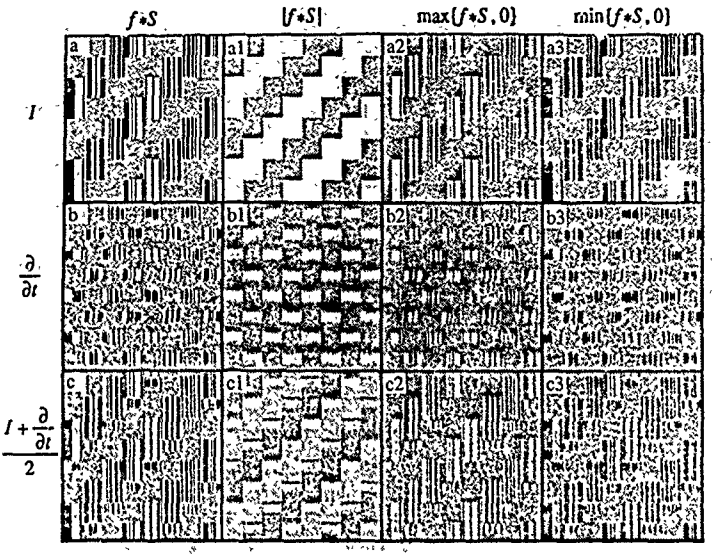


Figure 3: *Transformations of the contrast-modulating squarewave.* All 12 panels are *xt* cross-sections (with time running downward) of various transformations of stimulus 3.1, the *contrast-modulating squarewave*. Stimulus 3.1 itself is cross-sectioned in (a). The horizontal dimension is x, the vertical dimension is t with time increasing downward, and the stimulus is unvarying in y. The problem of perceiving leftward motion in the dynamic display whose *xt* cross section is represented by panel (a) is equivalent to the texture problem of perceiving orientation slanting down and to the left in the panel (a) itself. In the left-hand column are displayed cross-sections of 3 linear transformations of the stimulus: (a) the identity, (b) the partial derivative with respect to time, and (c) the average of the operators applied in (a) and (b). The next column (a1, b1, c1) shows the result of full-wave rectification (absolute value) of the corresponding (same-row) linear transformations; e.g., (a1) shows the result of full-wave rectifying the untransformed stimulus 3.1. Column three shows the positive half-wave components of the same-row linear transformations in column 1; column 4 shows the negative half-wave components. The functions in column 1 (linear transformations of the contrast-modulating squarewave) are all microbalanced; hence, the right-to-left motion displayed by the stimulus cannot be obtained from these transformations by standard motion analysis. Temporal differentiation (the second-row transformations) yields motion-ambiguous functions; rows 1 and 3 yield functions whose motion is extractable by standard motion analysis.

conditions, despite the fact that (as is easily proven from propositions 2.4.5a and 2.4.6) it is microbalanced.

Simple rectification exposes the motion of the amplitude-modulating squarewave. As suggested by Figs. 3a1, 3a2, and 3a3, simple full-wave or half-wave rectification (i.e. setting $f*$ to the identity in Eq. (1)) suffices to expose motion carried by amplitude-modulation. However, simple rectification fails to expose the motion in the following stimulus.

3.2. Stimulus: The contrast-reversing squarewave. A sideways stepping squarewave is used to alternately multiply the *contrast* of spatially independent noise by +1 and −1. Fig. 4a shows an *xt* cross-section of a squarewave that steps leftward 1/4 spatial-cycle at regular temporal intervals, reversing the *contrast* of black/white vertical bars as it moves. Like the amplitude-modulating squarewave, this contrast-reversing squarewave displays vivid leftward motion to all viewers under a broad range of viewing conditions; nonetheless, it is microbalanced (another easy consequence of propositions 2.4.5a and 2.4.6).

Simple rectification fails to reveal the motion of the contrast-reversing squarewave. As illustrated in Figs. 4a1, 4a2 and 4a3, simple rectification does not expose the motion of the contrast-reversing squarewave to standard motion analysis; full-wave rectification yields a uniform field, while half-wave rectification yields a mere dc-shifted rescaling of the original stimulus. Indeed, any purely spatial filter followed by rectification is equally ineffective at revealing this motion [27]

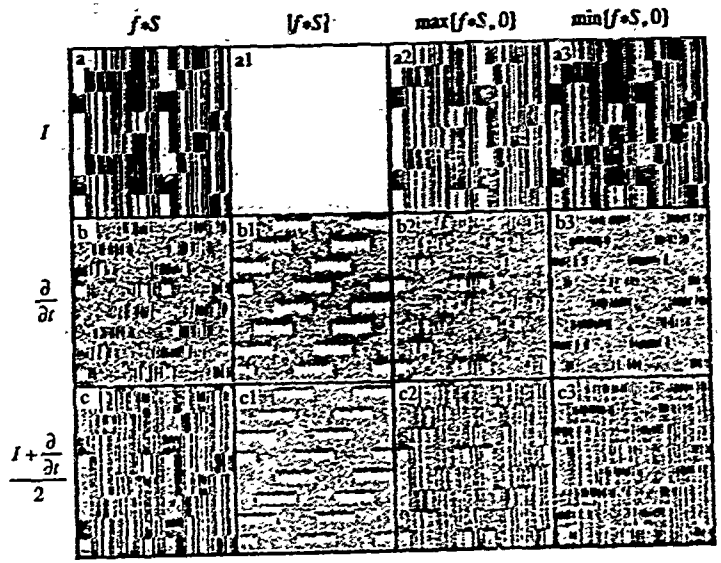


Figure 4. *Transformations of the contrast-reversing squarewave.* All 12 panels are *xt* cross-sections (with time running downward) of various transformations of stimulus 3.2, the *contrast-reversing squarewave*, which is itself is cross-sectioned in (a). See caption of Fig. 3 for a description of the transformations and panel arrangement. All of the functions in the left column (the linear transformations of the contrast-reversing squarewave) are microbalanced, so the right-to-left motion displayed by the stimulus cannot be obtained from these transformations by standard motion analysis. The purely pointwise transformations (the rectifications shown in the first row) also yield microbalanced functions and hence no simply-accessible motion. However, after any of the rectifying transformations in rows (b) or (c), the stimulus motion is accessible to standard motion analysis.

Temporal differentiation followed by rectification reveals the motion of the contrast-reversing squarewave. The obvious transformation to expose the motion of this stimulus to standard motion analysis is *temporal differentiation* followed by half-wave or full-wave *rectification*. The result of differentiating the contrast-reversing squarewave with respect to time is shown in Fig. 4b. The motion of this temporal derivative remains microbalanced (a consequence of propositions 2.4.4 II. and 2.4.5a). However, as suggested by Figs. 4b1, 4b2 and 4b3, either full-wave (Fig. 4b1) or half-wave (Figs. 4b2 & 4b3) rectification suffices to reveal the motion of the temporal derivative of the contrast-reversing squarewave to standard analysis. However,

Temporal differentiation followed by rectification fails to expose the motion of the amplitude-modulating squarewave. Differentiating the *amplitude-modulating* squarewave (Fig. 3a) with respect to time *sacrifices* all the motion content of this stimulus (See. Fig. 3b). The differentiated stimulus (Fig. 3b) is completely ambiguous in motion-content, and subsequent transformations (e.g. full- or half-wave rectification: Figs. 3b1, 3b2, 3b3) cannot reclaim the original stimulus motion.

To recapitulate: The motion of the amplitude-modulating squarewave (Fig. 3a) is exposed by simple half-wave or full-wave rectification (Figs. 3a1, 3a2, 3a3). However, rectification fails to expose the motion of the contrast-reversing squarewave (signal in Fig. 4a; rectifications in Figs. 4a1, 4a2, 4a3). On the other hand, temporal differentiation followed by half-wave or full-wave rectification suffices to expose the motion of the contrast-reversing squarewave to standard analysis (Figs. 4b, 4b1, 4b2, 4b3), but fails to reveal the motion of the amplitude-modulating squarewave (Figs. 3b, 3b1, 3b2, 3b3).

A single transformation which reveals the motion of both stimuli 3.1 and 3.2 to standard motion-analysis can easily be obtained by letting $f*$ of Eq. (1) be a temporal linear filter (spatial component = identity) with impulse response given by Fig. 5

The result of applying such a filter to the contrast-modulating squarewave is shown in Fig. 3c. As Figs. 3c1, 3c2, and 3c3 suggest, full- or half-wave rectification of the output (Fig. 3c) exposes the motion of the contrast-modulating square to standard analysis. And as Figs. 4c, 4c1, 4c2 and 4c3 indicate, the same transformations expose the motion of the contrast-reversing squarewave to standard analysis.

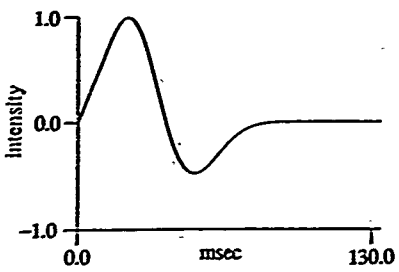


Figure 5: The impulse response of a temporal filter suitable to extract the motion of the contrast-modulating squarewave (Fig. 3a), and of the contrast-reversing squarewave (Fig. 4a). The filtered responses are shown in Figs. 3c and 4c. Subsequent rectification makes the motion accessible to standard motion analysis (see bottom row of Figs. 3 and 4).

4. Motion carried purely by spatial texture.

There are many microbalanced random stimuli whose motion depends on the spatiotemporal modulation of *spatial texture*. The most obvious transformations to expose texturally conveyed motion to standard motion analysis are given by $T(I) = r \circ (f * I)$, with the separable filter $f\circ$ being purely spatial (temporal component = identity). The spatial filter $f\circ$ should be viewed as a "texture-grabber". $f\circ$ will respond with varying power throughout regions of the visual field, depending on whether or not the texture to which it is tuned populates those regions. However, the output of a linear filter to a texture is positive or negative according to the phase of the texture. That is, multiplying the contrast of the texture by −1 will multiply the filter's output by −1. The purpose of rectification is to report the presence or absence of texture, independent of phase. The result $T(I)$ is a spatiotemporal function whose value reflects the movement of the $(f\circ)$-texture across the visual field as a function of time. Elaborations of this scheme have been applied to modeling texture perception by Caelli [32], Bergen & Adelson [33], and Sutter, Beck & Graham [34].

To study texturally conveyed motion, it is important to bypass not only first-order motion mechanisms, but also irrelevant second-order mechanisms, such as the temporal mechanisms proposed above for accessing the motion of the amplitude- and contrast-reversing squarewaves—stimuli 3.1 and 3.2). A particular subclass of microbalanced random stimuli serves this purpose.

4.1. Random stimuli microbalanced under all pointwise transformations.

Many signal transformations encountered in perceptual models can be expressed as cascades of pointwise $(r\circ)$ and space/time separable LSI transformations $(f\circ)$. For visual processing that is limited to such cascades, the following question is of considerable interest: What conditions must be satisfied by a random stimulus I in order that $r\circ I$ be microbalanced for *any* function $r: \mathbb{R} \to \mathbb{R}$? For any such I, the cascade $f\circ(r\circ I)$ does not suffice to reveal I's motion to standard analysis, as each successive transformation leaves the stimulus microbalanced. Thus I's motion can only be perceived by a mechanism which applies a cascade that includes a nontrivial LSI transformation followed by a pointwise nonlinearity.

Indeed, we have already encountered a simple example of such a stimulus, the contrast-reversing squarewave stimulus 3.2, which we shall call J in this discussion. We demonstrated above that simple rectification cannot expose the motion of J to standard analysis, and it is easy to see that this observation generalizes beyond rectifiers to all pointwise transformations: Any pointwise transformation applied to J yields a rescaled version of J plus a constant. Input and output are both microbalanced.

The motion carried by the contrast-reversing squarewave J is exposed by rectifying the output of certain LSI transformations (e.g. the temporal derivative) of J. However, no pointwise transformation applied by itself to J suffices to expose J's motion content to standard analysis. In studying the processing stages that mediate second-order motion perception, it may be of some importance to know that a given stimulus is "immune" to a certain transformation or a certain type of transformation (as J is immune to pointwise transformations). This motivates the following notion [27]:

4.1.1. Call any random stimulus I microbalanced *under* a given transformation T iff $T(I)$ is microbalanced.

In connection with pointwise transformations, we have the following two results [27]:

4.1.2. *Let I be a random stimulus such that, for any $(x,y,t),(x',y',t') \in \mathbb{Z}^3$, $I[x,y,t]$ and $I[x',y',t']$ have a continuous joint density. Then the following conditions are equivalent:*

1. *I is microbalanced under all pointwise transformations.*

2. *The joint density f of $I[x,y,t]$ with $I[x',y',t']$ and the joint density g of $I[x,y,t']$ with $I[x',y',t]$ satisfy*

$$f(p,q) + f(q,p) = g(p,q) + g(q,p)$$

for all $p, q \in \mathbb{R}$.

4.1.3. (Corollary) *For any random stimulus I, if the joint density of $I[x,y,t]$ with $I[x',y',t']$ is identical either to the joint density of $I[x,y,t']$ with $I[x',y',t]$ or to the joint density of $I[x',y',t]$ with $I[x,y,t']$ for every $(x,y,t),(x',y',t') \in \mathbb{Z}$, then I is microbalanced under all pointwise transformations.*

4.2. Texture quilts.

The results of section 4.1 can easily be applied to construct a wide variety of stimuli for which the first effective stage of processing for motion involves a non-pointwise

transformation. If, as most models presume, processing channels are restricted to cascaded pointwise and LSI transformations, then this initial transformation must be (non-trivially) LSI. By themselves, however, LSI operators are insufficient to expose the motion of microbalanced random stimuli to standard analysis.

Thus, interposed between the initial LSI transformation and standard motion analysis there must be a (nonlinear) pointwise transformation. For the contrast-reversing squarewave, the LSI transformation of temporal differentiation followed by the pointwise transformation of full-wave rectification suffices to expose the motion to standard analysis.

Like the contrast-reversing squarewave, the following stimuli (i) are microbalanced under all pointwise transformations, and (ii) display consistent apparent motion across independent realizations. Unlike the contrast-reversing squarewave, the texture-grabbing filters appropriate for the following stimuli are spatial rather than temporal. In fact, it can be shown [27] that each of the stimuli I presented in this section is microbalanced *under all purely temporal transformations*; i.e., under all transformations whose output at a given point (x, y, t) in space/time depends only on the history of input at the spatial point (x, y). Thus, none of the transformations that sufficed to expose the motion of the amplitude-modulating and contrast-reversing squarewaves would reveal the motion of I to standard motion analysis.

All the examples of this section exploit the same essential trick: briefly displayed patches of static, random-phased texture occur in specific spatiotemporal relations to each other, and appropriate measures are taken to ensure that the resulting stimulus is microbalanced under all pointwise transformations. We call such stimuli *texture quilts*. The texture quilts constructed in our examples (exemplars are shown in Figs. 6b, 7d, 8b and 8c) all display decisive apparent motion from left to right, when viewed either monocularly or binocularly from a distance such that they span about 4 horizontal retinal degrees, with frames displayed at 15 Hz.

Binary texture quilts.

The easiest constructions of quilts that are microbalanced under all purely temporal transformations use stimuli that have only two contrast values. We show how to construct a generic binary-valued quilt and provide some specific examples.

4.2.1. A general technique for constructing binary texture quilts that are microbalanced under all purely temporal transformations. Let $\alpha \subset Z^2$ be a set of points in space (those which will take nonzero values at some time during the display). For the number N of frames comprising the quilt, associate with frames 1 through N a family

$$\phi_1, \phi_2, \ldots, \phi_N$$

of jointly independent random variables, each of which takes the value 1 or -1 with equal probability. In addition, associate with frames 1 through N, a family

$$f_i, \quad i = 1, 2, \ldots, N$$

of functions, with f_i assigning 0 throughout all frames except the i^{th}, and within frame i, assigning 0 everywhere except α, with α being mapped into $\{1, -1\}$. Then, construct the stimulus

$$B = \phi_1 f_1 + \phi_2 f_2 + \ldots + \phi_N f_N.$$

It is easily derived from corollary 4.1.3 that B is microbalanced under all pointwise transformations. The proof that B is microbalanced under all purely temporal transformations is in [27].

4.2.2. Stimulus: The sidestepping, randomly contrast-reversing, vertical edge. Figure 6b displays nine frames comprising a particularly simple binary texture quilt. Note that the vertical dimension of Fig. 6b combines time and vertical space. The representation in Fig. 6b is precisely equivalent to a strip of movie film with frames arranged vertically above each other, separated by grey lines. Between successive grey lines is displayed the actual two-dimensional luminance function displayed to subjects. Fig. 6a shows the functions f_1 through f_9 used in the construction. f_1 assigns the value −1 to all points (x, y, t) within the spatiotemporal block of the first frame, and 0 to all other points. f_2 assigns the value 1 to the points in the leftmost eighth of the second frame, the value −1 to the points in the right seven eighths, and 0 to all points outside the second frame. The functions coloring successive frames shift the bright/dark edge rightward through the frame until in frame 9, the field is uniformly bright. Multiplying each frame $i = 1, 2, \ldots, 9$ by its associated random variable ϕ_i yields, in this particular realization, the stimulus given in Fig. 6b.

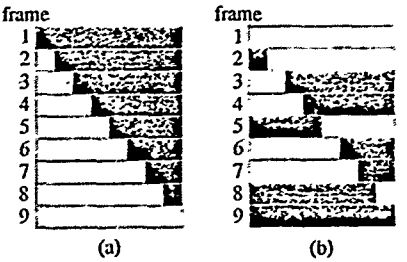


Figure 6: *Edge-driven motion from an ordinary edge and from a binary texture quilt.* (a) A rightward moving light/dark edge visible to first- and second-order motion detectors. Nine frames are shown; each frame shows exactly what is displayed, an area of contrast +1 and area of contrast −1. (b) A realization of the *sidestepping, randomly contrast-reversing vertical edge*. This random stimulus is microbalanced under all purely temporal transformations; therefore its rightward motion remains inaccessible to standard motion analysis even after an arbitrary, purely temporal transformation. Each of the frames 1 – 9 of (b) was derived from the corresponding frame of (a) by multiplying that entire frame by a random variable that takes the value 1 or −1 with equal probability. The nine frame random variables are jointly independent.

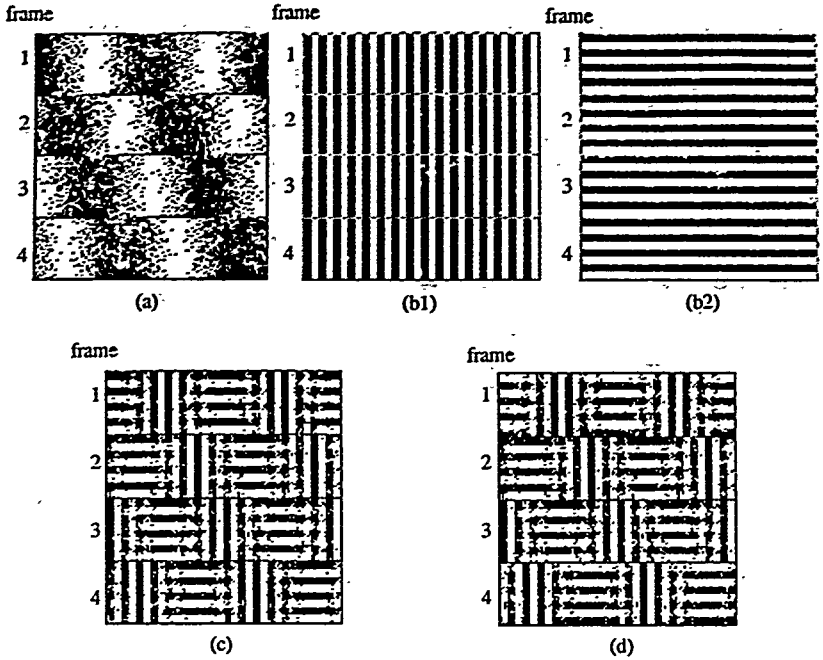


Figure 7. *Orientation-driven second-order motion from a binary texture quilt.* (a) Four frames of a probabilistically defined sinewave grating that steps rightward 90 degrees between frames. The rightward motion in (a) is accessible to all motion detectors. (b1) Four frames of a static, vertical squarewave grating; (b2) Four frames of a static horizontal squarewave grating. (c) A rightward translating texture pattern. For every white point in (a), the corresponding value in (c) is chosen from the vertical square-wave grating in (b1); for every black point, the corresponding value in (c) is chosen from the horizontal square-wave grating in (b2). Stimulus (c) is not microbalanced, its motion is accessible to standard motion analysis. (d) A texture quilt. The frames of (d) are derived by multiplying the corresponding frames of (c) by jointly independent random variables, each of which takes the value 1 or −1 with equal probability. The texture quilt realized in (c) is microbalanced under all purely temporal transformations, therefore its rightward motion is unavailable to standard motion analysis, even after an arbitrary, purely temporal transformation.

The motion displayed by this quilt is clearly driven by the randomly contrast-reversing edge that steps from left to right through the course of the display. Almost any bandpass spatial filter followed by a rectifier will suffice to expose this motion to standard analysis. The following quilt requires a more specifically tuned texture-grabbing spatial filter.

4.2.3. Stimulus: Oppositely oriented, randomly contrast-reversing squarewaves selected by a drifting grating. In Fig. 7d are displayed the four frames comprising another binary texture quilt also constructed using technique 4.2.1. Figure 7c shows the functions f_1, f_2, f_3, and f_4 used in the construction. Each of these frames was constructed by using the corresponding frame of the probabilistically defined, rightward stepping sinusoid of Fig. 7a to sample between the two squarewave gratings shown in Figs. 7b1 and 7b2. The texture quilt realized in Fig. 7d is derived by randomly reversing the contrast of each of the frames of Fig 7c. For the realization given in Fig. 7d, the random variables ϕ_1, ϕ_2, ϕ_3 and ϕ_4 used to multiply the frames of Fig. 7c take the values −1, −1, 1, and 1 respectively.

Sinusoidal texture quilts.

It is simple to elaborate technique 4.2.1 to a method for constructing quilts involving textures of arbitrarily many contrast values. We illustrate the principle in the construction of a generic quilt comprised of patches of sinusoidal grating and we provide two specific examples.

4.2.4. A general technique for constructing sinusoidal texture quilts microbalanced under all purely temporal transformations. A generic sinusoidal quilt has N frames. Pixels of each frame are filled by choosing between a pair of sinusoids assigned to that frame. The critical constraints (to insure that the resulting stimulus will be microbalanced under all purely temporal transformations) are that the different sinusoids thus patched together, within a given frame and across different frames, must be of equal amplitude and have jointly independent, uniformly distributed random phases.

Specifically, for $i = 1, 2, \ldots, N$, with N the number of frames comprising the quilt, let W_i be a function, temporally constant within frame i, assigning either 1 or -1 to all points (x, y, t) in the i^{th} frame, and 0 to all points outside the i^{th} frame. We use W_i to sample between static sinusoidal gratings with random phases and different spatial frequencies. Apparent motion can often be generated with such displays by shifting each successive sampling function W_{i+1} in a fixed direction relative to W_i.

Let

$$\omega_1, \theta_1, \tilde{\omega}_1, \tilde{\theta}_1, \omega_2, \theta_2, \tilde{\omega}_2, \tilde{\theta}_2, \ldots, \omega_N, \theta_N, \tilde{\omega}_N, \tilde{\theta}_N$$

be integers. For each frame i of the texture quilt being constructed we shall use W_i to sample between two sinusoids, C_i and $\tilde{C}_i$. For some integer P (independent of frame), C_i has a spatial frequency of ω_i/P cycles per horizontal pixel and θ_i/P cycles per vertical pixel. $\tilde{C}_i$ has a spatial frequency of $\tilde{\omega}_i/P$ cycles per horizontal pixel and $\tilde{\theta}_i/P$ cycles per vertical pixel.

The phases of all of the sinusoids patched together in the quilt are independent random variables. To be precise, let

$$\rho_1, \tilde{\rho}_1, \rho_2, \tilde{\rho}_2, \ldots, \rho_N, \tilde{\rho}_N,$$

be jointly independent random variables, each assuming with equal probability a value from amongst the integers $0, 1, \ldots, P-1$. Then for all $(x, y, t) \in Z^3$ set

$$S = \sum_{i=1}^{N} S_i,$$

where, for each i,

$$S_i[x,y,t] = \begin{cases} \cos(2\pi(\omega_i x + \theta_i y + \rho_i)/P) & \text{if } W_i[x,y,t] = 1, \\ \cos(2\pi(\tilde{\omega}_i x + \tilde{\theta}_i y + \tilde{\rho}_i)/P) & \text{if } W_i[x,y,t] = -1, \\ 0 & \textit{otherwise.} \end{cases}$$

Like the generic binary texture quilt B, S is microbalanced under all purely temporal transformations [27].

4.2.5. Stimulus: Oppositely oriented, random-phased sinusoids selected by a drifting grating. The sinusoidal analog to the binary texture quilt of Fig. 7d is shown in Fig. 8b. In Fig. 8a are shown the functions W_1, W_2, W_3, and W_4 used to select between horizontal and vertical gratings. For this quilt, $\tilde{\omega}_i = \theta_i = 0$, for $i = 1, 2, 3, 4$; and for some integer F (with F/P the number of cycles per pixel), $\omega_i = \tilde{\theta}_i = F$.

The motion displayed by the texture quilt of Fig 8b evidently depends on the difference in orientation between the textures mixed in each frame. Of course, we can just as easily keep orientation constant and vary spatial frequency instead.

4.2.6. Stimulus: Random-phased sinusoids of different spatial frequencies, selected by a drifting grating. Figure 8c shows a texture quilt using the sampling functions of Fig. 8a, but setting $\omega_i = \theta_i = 2\tilde{\omega}_i = 2\tilde{\theta}_i$ for $i = 1, 2, \ldots, 4$.

The empirical observations with texture quilts are that motion can be perceived when texture patches move across the field, even when the texture-conveyed motion is contrived so that there are no spatiotemporal correlations in the stimulus to support standard motion analysis [11,17], and when second-order temporal processing can be excluded [27]. These texture-conveyed motions are detected by convolving the input stimulus with a spatial texture-grabbing filter tuned to the moving texture, then rectifying the output of the filter (to indicate the presence or absence of the texture), and subjecting the rectified output to standard motion analysis. That supraordinate texture orientation is easily perceived in the x,y representations of the texture-conveyed motion (Figs. 7d, 8b and 8c) indicates that there exists second-order orientation processing of textures in the space domain analogous to the second-order motion processing of textures in the motion domain.

5. Summary.

Section 1 introduced the distinction between first- and second-order motion mechanisms. Section 2 reviewed the fundamental results concerning drift-balanced and microbalanced random stimuli. Microbalanced random stimuli are useful in the study of second-order motion perception because (i) they are guaranteed *not to systematically stimulate* first-order (Fourier-energy analytic or autocorrelational) motion mechanisms, and (ii) it is easy to produce microbalanced random stimuli that display consistent, compelling apparent motion across independent realizations.

Section 3 described microbalanced random stimuli that displayed different types of apparent motion. The contrast-modulating squarewave (Stimulus 3.1) suggests that some instances of microbalanced motion may be exposed to standard motion analysis by simple rectification. The contrast-reversing squarewave (stimulus 3.2) suggests that other instances of microbalanced motion are exposed by rectifying the temporal derivative of the stimulus. Moreover, the motion of stimulus 3.1 can *not* be exposed by temporal differentiation followed by rectification, whereas the motion of stimulus 3.2 can *not* be exposed by simple rectification. A temporal filter with the impulse response given in Fig. 5 (including terms for both temporal differentiation and temporal identity), followed by rectification, does suffice to expose the motion of both stimuli 3.1 and 3.2 to standard motion analysis. For each of these stimuli, the optimal *spatial* filter to expose the motion is the identity.

Section 4 introduced the notion of a random stimulus *microbalanced under all pointwise transformations*. Section 4.1 provided necessary and sufficient conditions for a random stimulus to be of this sort. Such stimuli I are significant because pointwise transformations applied directly to I merely result again in microbalanced random functions, thus the first

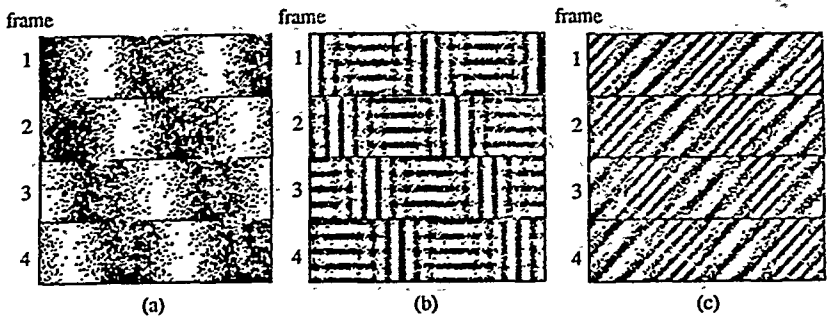


Figure 8: *Sinusoidal texture quilts—motion driven by differences in orientation or in spatial frequency.* The 4 frames in (a) are used to select between two sinusoidal patterns. Stimuli (b) and (c) are realizations of two such random stimuli, each of which is microbalanced under all purely temporal transformations. The sinusoids mixed in (b) differ in orientation, while the sinusoids mixed in (c) have the same orientation, but differ in spatial frequency. The phases of sinusoids are jointly independent across frames, and across sinusoids of different frequency mixed in the same frame.

transformation in any respect effective at exposing I's motion to analysis must be non-pointwise. If the transformations applied to the visual signal are limited to cascades of (i) linear shift-invariant operators and (ii) pointwise operators, then the first processing stage effective in revealing the motion of I must be a nontrivial linear transformation. Moreover, since I is microbalanced, this linear filter must be followed by at least a pointwise nonlinearity for I's motion to be revealed to standard analysis.

Section 4.2 illustrated random stimuli—*texture quilts* (stimuli 4.2.2, 4.2.3, 4.2.5 and 4.2.6)—that yielded compelling texture-conveyed apparent motion. These stimuli were microbalanced under all purely temporal transformations. Their motion cannot be exposed by simple rectification, nor indeed by any *purely temporal* transformations, no matter how nonlinear. The perception of texture quilt motion can be modeled in terms of a *spatial* texture-grabbing filter followed by rectification and standard motion analysis. Thus, the minimal system to account for all the demonstrations of second-order motion perception presented here would consist of a temporal filter that has both an identity and a temporal differentiation component, a band-selective spatial filter followed by a rectifier and standard motion analysis.

Acknowledgements.

The research reported here was supported by USAF Life Science Directorate, Visual Information Processing Program, Grants 85-0364 and 88-0140.

References

[1] Adelson, E.H. and J. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Am. A*, 2, 2, 284-299, 1985.

[2] Watson, A B. and A.J. Ahumada, "A model of human visual motion sensing," *J. Opt. Soc. Am. A*, 2, 2, 322-342, 1985.

[3] van Santen, J P.H. and G. Sperling, "A Temporal Covariance Model of Motion Perception," *J. Opt. Soc. Am. A*, 1, 451-473, 1984.

[4] van Santen, J.P H. and G. Sperling, "Elaborated Reichardt Detectors," *J. Opt. Soc. Am. A*, 2, 2, 300-321, 1985.

[5] Heeger, D.J., "A model for the extraction of image flow," *J Opt. Soc Am. A*, 4, 8, 1455-1471, 1987.

[6] Watson, A B. and A.J. Ahumada, "A Look at Motion in the Frequency Domain," NASA Technical Memorandum 84352, 1983.

[7] Marr, D. and S. Ullman, "Directional selectivity and its use in early visual processing," *Proc. R. Soc. Lond B.*, 211, 151-180.

[8] Marr, D. *Vision*, W. H. Freeman & Co., 1982.

[9] Anandan, P. "A unified perspective on computational techniques for the measurement of visual motion," *DARPA IU workshop proc.*, 1987.

[10] Waxman, A. M. and F. Bergholm, "Convected activation profiles and image flow extraction," *Laboratory for Sensory Robotics Technical Report 4*, College of Engineering, Boston U., 1987.

[11] Ramachandran, V.S., V.M. Rao and T.R. Vidyasagar, "Apparent movement with subjective contours," *Vision Res*, 13, 1399-1401, 1973

[12] Sperling, G., "Movement perception in computer-driven visual displays," *Behavior Research Methods and Instrumentation*, 8, 144-151, 1976.

[13] Petersik, J T., K.I. Hicks and A Pantle, "Apparent movement of successively generated subjective figures," *Perception*, 7, 371-383, 1978.

[14] Lelkens, A.M.M. & J.J. Koenderink, "Illusory motion in visual displays," *Vision Res*, *24*, 1083-1090, 1984.

[15] Cavanagh, P., J. Boeglin and O.E. Favreau, "Perception of motion in equiluminous kinematograms," *Perception*, 14, 151-162, 1985

[16] Derrington, A.M. and D.R. Badcock, "Separate detectors for simple and complex grating patterns?" *Vision Res*, 25, 1869 1878, 1985

[17] Green, M., "What determines correspondence strength in apparent motion," *Vision Res*, 26, 599-607, 1986.

[18] Prazdny, K., "What variables control (long range) apparent motion?" *Perception*, 15, 37-40, 1986.

[19] Pantle, A. and K Turano, "Direct comparisons of apparent motions produced with luminance, contrast modulated (CM), and texture gratings" *Investigative Ophthalmology and Visual Science*, 27, 3, 1986.

[20] Derrington, A.M. and G.B. Henning, "Errors in direction-of-motion discrimination with complex stimuli," *Vision Res*, 27, 61-75, 1987.

[21] Turano, K. and A. Pantle, "On the mechanism that encodes the movement of contrast variations - I: velocity discrimination," submitted.

[22] Bowne, S B., S P. McKee, & D A. Glaser, "Motion interference in speed discrimination," in press.

[23] Cavanagh, P., Arguin, M. and M. von Grunau, "Inter-attribute apparent motion," in press.

[24] Chubb, C. and G. Sperling, "Drift-balanced random stimuli: A general basis for studying non-Fourier motion perception," *Investigative Ophthalmology and Visual Science*, 28, p. 233, 1987.

[25] Chubb, C. and G. Sperling, "Processing stages in non-Fourier motion perception," *Investigative Ophthalmology and Visual Science*, 28, p. 266, 1988.

[26] Chubb, C. and G. Sperling, "Drift-balanced random stimuli: A general basis for studying non-Fourier motion perception," *J. Opt. Soc. Am. A.*, 5, 11, 1986-2007, 1988.

[27] Chubb, C. and G. Sperling, "Texture quilts: basic tools for studying motion from texture," *Mathematical Studies in Perception and Cognition*, 88-1, New York University, Department of Psychology, 1988.

[28] Chubb, C. and G. Sperling, "Two motion-perception mechanisms revealed through distance-driven reversal of apparent motion," *Proc. Natl. Acad. Sci. USA*, 86, 1988, in press.

[29] Watson, A.B., A.J. Ahumada and J.E. Farrell, "The window of visibility: A psychophysical theory of fidelity in time-sampled motion displays," NASA Technical Paper 2211, 1983.

[30] Watson, A.B., A.J. Ahumada and J.E. Farrell, "The window of visibility: A psychophysical theory of fidelity in time-sampled motion displays," *J. Opt. Soc. Am. A*, 3, 3, 300-307, 1986.

[31] Reichardt, W., "Autocorrelation, a principle for the evaluation of sensory information by the central nervous system," in *Sensory Communication*, W. A. Rosenblith, ed., Wiley, New York, 1961.

[32] Caelli, T., "Three processing characteristics of visual texture segregation," *Spatial Vision*, 1, 1, 19-30, 1985.

[33] Bergen, J.R. and E.H. Adelson, "Early vision and texture perception," *Nature*, 333, 6171, 363-364, 1988.

[34] Sutter, A., J. Beck and N. Graham, "Contrast and spatial variables in texture segregation: testing a simple spatial frequency channels model," in press.

Drift-balanced random stimuli: a general basis for studying non-Fourier motion perception

Charles Chubb and George Sperling

Human Information Processing Laboratory, Psychology Department, New York University, 6 Washington Place, New York, New York 10003

Received November 17, 1987; accepted June 7, 1988

To some degree, all current models of visual motion-perception mechanisms depend on the power of the visual signal in various spatiotemporal-frequency bands. Here we show how to construct counterexamples. visual stimuli that are consistently perceived as obviously moving in a fixed direction yet for which Fourier-domain power analysis yields no systematic motion components in any given direction. We provide a general theoretical framework for investigating non-Fourier motion-perception mechanisms; central are the concepts of drift-balanced and microbalanced random stimuli. A random stimulus S is drift balanced if its expected power in the frequency domain is symmetric with respect to temporal frequency, that is, if the expected power in S of every drifting sinusoidal component is equal to the expected power of the sinusoid of the same spatial frequency, drifting at the same rate in the opposite direction. Additionally, S is microbalanced if the result WS of windowing S by any space-time-separable function W is drift balanced. We prove that (i) any space-time-separable random (or nonrandom) stimulus is microbalanced; (ii) any linear combination of pairwise independent microbalanced (respectively, drift-balanced) random stimuli is microbalanced and drift balanced if the expectation of each component is uniformly zero; (iii) the convolution of independent microbalanced and drift-balanced random stimuli is microbalanced and drift balanced; (iv) the product of independent microbalanced random stimuli is microbalanced; and (v) the expected response of any Reichardt detector to any microbalanced random stimulus is zero at every instant in time. Examples are provided of classes of microbalanced random stimuli that display consistent and compelling motion in one direction All the results and examples from the domain of motion perception are transposable to the space-domain problem of detecting orientation in a texture pattern.

1. INTRODUCTION

Central to the study of human visual motion perception is the relationship between perceived motion and the Fourier transform of the spatiotemporal visual stimulus Points in the domain of the spatiotemporal Fourier transform correspond to drifting sinusoidal gratings. For a wide range of spatial and temporal frequencies, such drifting sinusoids are perceived to move uniformly across the visual field, and their apparent speed and direction are direct functions of spatiotemporal frequency. For the most part, the motion displayed by simple linear combinations of such gratings reflects quite reasonably the individual contributions of the components.[1,2]

Indeed, current models of human motion perception implicitly or explicitly involve some degree of Fourier decomposition (bandpass filtering) of the image stream.[1-6] Generally, of course, the decomposition is localized to finite temporal intervals and subregions of the visual field.

It has long been realized, however, that certain sorts of apparent motion cannot be understood directly in terms of their power spectra.[7-14] For instance, much attention has been focused on sums of drifting gratings of slightly different, high spatial frequencies.[10-12] In general, the perceived velocity of such stimuli is determined not directly by the frequencies of the summed components but by the pattern of beats at their difference frequency.

Sperling[13] demonstrated "movement without correlation" in a different stimulus whose Fourier transform, when computed globally or locally, contained no consistent moving components and yet was perceived to move decisively in a fixed direction. Subsequently, Petersik *et al.*[14] studied similar displays in an effort to clarify the relationship between stage 1 (autocorrelational, Fourier) mechanisms and the higher-order stage 2 mechanisms mediating the perception of what we call[15] non-Fourier motion.

The purpose of this paper is to provide (i) a general theoretical basis and (ii) an array of specific tools for studying non-Fourier motion-perception mechanisms.[16]

2. ANALYZING A STIMULUS: INTUITIVE FOURIER DECOMPOSITION

We begin with a brief, informal discussion to show how particular motion stimuli can be analyzed into drifting sinusoids For illustration we use one-dimensional spatiotemporal stimuli that move either to the left or to the right and whose luminance varies in only the horizontal dimension, although all the results that we derive apply in all cases to stimuli of two spatial dimensions and time. A one-dimensional, horizontally moving stimulus is represented conveniently by a two-dimensional function $l(x, t)$, where x (the horizontal axis) indicates the spatial pattern of luminance and t (the vertical axis, with time increasing upward) indicates the temporal luminance pattern. In this representation, usually it is immediately obvious which way the dominant Fourier components of l tend to slope (up and to the left

0740-3232/88/111986-22$02 00

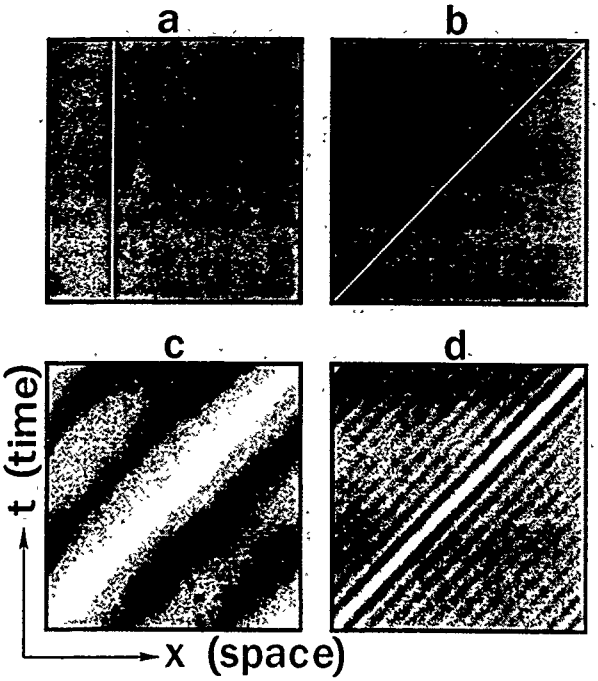


Fig 1. Spatiotemporal Fourier analysis of a rightward stepping bar The abscissa represents horizontal space, the ordinate represents time. a, One frame of a movie of a rightward stepping vertical bar. b, Horizontal-temporal cross section of a rightward-stepping vertical bar. c, Approximation to the rightward-stepping bar obtained by taking an equally weighted sum of $\{\cos(2\pi n(x/X - t/T))|n = 1, 2\}$. d, Approximation to the rightward-stepping bar obtained by taking an equally weighted sum of $\{\cos(2\pi n(x/X - t/T))|n = 1, 2, \ldots, 12\}$.

or up and to the right). For example, Fig. 1a represents a single frame of a white vertical bar, extended up and down through the field of vision. Figure 1b shows the space-time representation of the bar in Fig 1a, which appears at the left at time zero and moves at a constant rate to the right during the time course of the display.

For the moment, we shall generalize broadly, using the word sum to describe both finite and countable summations as well as integrations over bounded and unbounded real intervals. In this case, we can do approximate justice to some basic facts about visual stimuli and their Fourier transforms without getting bogged in technicalities Any spatiotemporal stimulus I can be decomposed into a weighted sum of appropriately phase-shifted, drifting sinusoidal gratings. Moreover, this sum is unique: that is, there is only one assignment of weights and phases to drifting gratings that recaptures I in the corresponding sum.

Indeed, the Fourier transform of I is often defined to be the function that makes this assignment. There are, however, various other commonly encountered equivalent definitions of the Fourier transform (one of which we shall shortly adopt) that may be more convenient for certain purposes.

Example: Fourier Components of a Rightward-Stepping Vertical White Bar

Most of the action of the moving bar stimulus I defined by Figs. 1a and 1b takes place along the line $L = \{(x, t)|x = t\}$ in Fig. 1b, that is, the points at which I deviates most from its mean value are along this line. For our purposes, the most useful indicator of where the action is in a given stimulus f is the squared deviation of f from its overall mean value at each point in its domair. As is clear, I deviates most energetically from its mean along the line L.

What spatiotemporal sinusoidal gratings are weighted most heavily in the Fourier sum yielding I? A good way to answer this question is to ask another What gratings can be shifted in phase so as to match I most closely? Those sinusoids that can be shifted so as to have high values where I has high values and low values where I has low values are the ones that will figure most heavily in the weighted sum com-

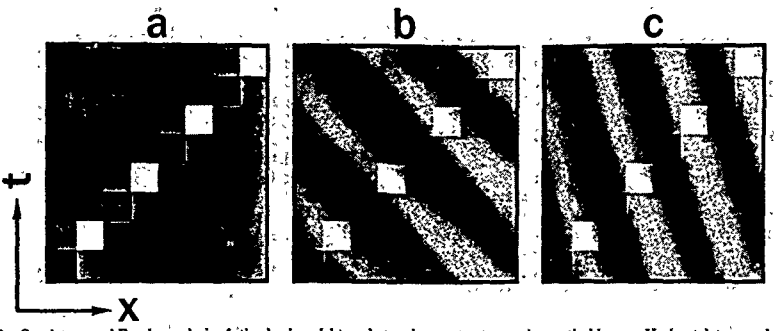


Fig 2. Spatiotemporal Fourier analysis of stimulus h, a rightward-stepping, contrast-reversing vertical bar. a, Horizontal-temporal cross section of h. b, Horizontal-temporal cross section of a vertical, leftward-drifting sinusoid, which correlates well with h: $\cos(2\pi(2x/X + 2t/T) - \pi/2)$. c, Horizontal-temporal cross section of a more slowly leftward-drifting sinusoid, which also correlates well with h: $\cos(2\pi(3x/Xz + t/T) - \pi/2)$.

posing l. In short, those gratings that can be phase shifted so as to correlate best with l will have the highest amplitudes (weights) in the sum.

The sinusoidal gratings that correlate best with $l(x, t)$ of Fig. 1b are those that assume the value 1 along the line L, that is, all the sinusoids in the set

$$\Omega = \{\cos(\alpha x - \alpha t) | \alpha \in \mathbb{R}\}.$$

Figures 1c and 1d illustrate how l is approximated more and more closely by taking sums involving more and more (equally weighted) elements of Ω.

Example 1: Rightward-Stepping, Contrast-Reversing Vertical Bar

Contrast-reversing stimuli are critical for understanding the implications of Fourier analysis. Note first that, as in the case of l defined in Fig. 1, most of the power of h in Fig. 2 is centered along the line L. However, the elements of Ω contribute no power to h. To see this, note that the value of h flipflops around the mean luminance along L, while the value of any element $C \in \Omega$ remains constant; thus the value of the product of h with C will flipflop (with h) around the mean luminance over the points of L and will be zero everywhere else. Consequently, the sum taken over all points (x, t) of the product $h(x, t)C(x, t)$ is zero. This is equivalent to saying that the correlation of h with C is zero.

On the other hand, the function

$$C(x, t) = \cos(\alpha x + \beta t + \rho)$$

correlates positively with h when α and β are chosen so that the crests and troughs of C slope across L and oscillate at an appropriate frequency. ρ can then be chosen to lay the crests of C across the bright regions of h and the troughs across the dark regions. Examples of sinusoids that correlate well with h are given in Figs. 2b $[\cos(3x + t - \pi/2)]$ and 2c $[\cos(2x + 2t - \pi/2)]$.

Direction of Drift in Sinusoidal Gratings

For each nonnegative real number α, $\cos(\alpha x - \alpha t)$ drifts from left to right. By contrast, $\cos(\alpha x + \alpha t)$ drifts at the same rate from right to left. For any $\omega, \tau, \rho \in \mathbb{R}$, if $\omega = 0$, the grating

$$C(x, t) = \cos(\omega x + \tau t + \rho)$$

has constant value over space but oscillates in time with frequency τ. Otherwise (if $\omega \neq 0$) C drifts with speed $|\tau/\omega|$; it drifts rightward if $\tau/\omega \leq 0$ and leftward if $\tau/\omega > 0$. Accordingly, we call C rightward drifting if $\tau/\omega < 0$, leftward drifting if $\tau/\omega > 0$, and stationary if $\tau = 0$.

3. THE MOTION-FROM-FOURIER-COMPONENTS PRINCIPLE

For any real-valued function, f, the sum (taken over all points in the domain of f) of the squared values of f is called the power in f. Parseval's relation states that the power in f is proportional to the sum of the squared amplitudes of the sinusoids into which f can be (uniquely) decomposed.

Thus, in particular, we can tally up the power in a dynamic visual stimulus either point by point in space–time or drifting sinusoid by drifting sinusoid. Of course, considering the unambiguous, uniform apparent motion displayed by drifting sinusoidal gratings, it would seem to make more sense for a motion-perception system to do its power accounting across the sinusoids composing the stimulus.

These considerations lead naturally to a commonly encountered general rule for predicting the apparent motion of an arbitrary horizontal stimulus $l(x, t)$: For l considered as a linear combination of sinusoidal gratings, compare the power in l of the rightward-drifting gratings with the power of the leftward-drifting gratings, if most of l's power is contributed by rightward-drifting gratings, then perceived motion should be to the right. If most of the power resides in the leftward-drifting gratings, perceived motion should be to the left. Otherwise l should manifest no decisive motion in either direction.

This prediction rule for horizontally moving stimuli is a restricted version of the motion-from-Fourier-components (MFFC) principle: More generally, let L be any visual stimulus; that is, $L: X \times Y \times T \rightarrow \mathbb{R}$, for bounded real intervals X, Y, and T, where for any $(x, y, t) \in X \times Y \times T$, $L(x, y, t)$ is

construed as the luminance of a point (x, y) in a visual field at time t. A more general version of the MFFC principle is as follows: For L to exhibit motion in a certain direction in the neighborhood of some point $(x, y, t) \in \mathbb{R}^3$, there must be some spatiotemporal volume Δ in some sense proximal to (x, y, t) such that the Fourier transform of L computed locally across Δ has substantial power over some regions of the frequency domain whose points correspond, in the space–time domain, to sinusoidal gratings whose direction of drift is consonant with the motion perceived.

That any standard version of the MFFC principle cannot account for all phenomena associated with human motion perception was demonstrated by Sperling,[13] who described the following three-flash stimulus. Frame 0 is a rectangular block of contiguous small squares, each of which is independently painted black or white with equal probability. In frame 1, a subblock B_1 of frame 0 is scrambled (that is, in frame 1, each component square within B_1 is independently repainted black or white with equal probability). In frame 2 a different subblock, B_2, is scrambled. For many sizes of rectangles and frame presentation rates, such a stimulus elicits apparent motion in the direction from B_1 to B_2; nonetheless, it is unlikely to correlate significantly with any given spatiotemporal sinusoidal grating.

It is our purpose here to build on these observations. We shall first give precise formulation to the notion of a random stimulus and then define a certain class of random stimuli (the class of drift-balanced random stimuli) that is useful in studying visual perception (since any motion displayed by a drift-balanced random stimulus cannot be explained in terms of the MFFC principle). We proceed to show that the (spatiotemporal) convolution of two drift-balanced random stimuli is drift balanced and mention some of the psychophysical implications of this fact. In proposition 3 below we prove that linear combinations of certain drift-balanced random stimuli are themselves drift balanced (this result, which is illustrated with a variety of stimulus examples, is particularly useful in constructing drift-balanced random stimuli that display consistent apparent motion across independent realizations). In Section 7 we provide an alternative characterization of the class of drift-balanced random stimuli in terms of simple point-delay Reichardt detectors (or autocorrelation coefficients) and apply this characterization to distinguish the subclass of drift-balanced random stimuli that we call microbalanced. A random stimulus I is microbalanced if, for any space–time-separable function W, the result WI of windowing I by W is drift balanced. We derive a collection of basic results about microbalanced random stimuli and show that, in fact, all the demonstration stimuli previously defined (demonstrations 1–5 below) are microbalanced. Among other things, we prove that the expected response of any elaborated Reichardt detector[1,2] to any microbalanced random stimulus is zero at any instant in time. Finally, we observe some salient psychophysical properties of microbalanced random stimuli and discuss some of the possible explanations of the non-Fourier motion elicited by such stimuli

4. PRELIMINARIES

In this paper we deal with properties of random stimuli Roughly speaking, a random stimulus is a jointly distributed family of random variables assigned to a grid of locations covering the visual field across time. In this section we collect the tools appropriate for dealing with such objects. This section is split into two subsections, one devoted to continuous random variables, in which we introduce explicitly some notation for handling integration and define a density, and one devoted to discrete dynamic visual stimuli and their Fourier transforms, in which we identify a stimulus [an assignment of luminance (nonnegative, real values) to a regular grid of points throughout visual space and time] with its contrast modulation function (the normalized deviation of luminance from its mean) and introduce frequency-domain notation.

Continuous Random Variables

Our stimuli are real-valued, randomly varying functions of a discrete domain. The luminances assigned to points (pixels) are, in general, jointly distributed random variables. The basic definitions and proofs that we present here presuppose that these random variables are real valued and continuous. (In general, the discrete-case analogs are simpler and should be obvious.)

Let $\mathbb{Z}$ ($\mathbb{Z}^+$) denote the set of integers (positive integers), and let $\mathbb{R}$ ($\mathbb{R}^+$) denote the real (positive real) numbers.

The following conventions are useful. As usual, call any subset $\alpha \subseteq \mathbb{R}$ an interval if and only if (iff), for any $x, z \in \alpha$ and any $y \in \mathbb{R}$, if $x \leq y \leq z$, then $y \in \alpha$; more generally, for any $k \in \mathbb{Z}^+$, call any subset $\beta \subseteq \mathbb{R}^k$ an interval of $\mathbb{R}^k$ iff β is the Cartesian product of (possibly unbounded) real intervals $\beta_0, \beta_1, \ldots, \beta_{k-1}$. In this case, for any function $f: \mathbb{R}^k \to \mathbb{R}$, it is convenient to indicate the integral of f over β, if it exists, as

$$\int_\beta f(\nu)\mathrm{d}\nu.$$

Moreover, we call any nonnegative, real-valued function f of $\mathbb{R}^k$ a density iff f is integrable over $\mathbb{R}^k$ and

$$\int_{\mathbb{R}^k} f(\nu)\mathrm{d}\nu = 1.$$

Discrete Dynamic Visual Stimuli and Their Fourier Transforms

Contrast Modulation

Luminance is physically constrained to be a nonnegative quantity. Psychophysically, however, the significant quantity is contrast, the normalized deviation at each time t of luminance at each point (x, y) in the visual field from a base level, or level of adaptation, which reflects the average luminance over points proximal to (x, y, t) in space and time. We shall restrict our attention throughout this paper to stimuli for which it may be assumed that the base luminance level μ is uniform over the significant spatiotemporal locations in the display. In practice, this condition is met if (i) subjects are adapted sufficiently to a field of uniform luminance μ before the onset of non-μ luminances and (ii) the duration over which non-μ luminances are displayed is sufficiently brief.

For any stimulus L with base luminance μ, call the function I satisfying

$$L = \mu(1 + I) \tag{1}$$

the *contrast modulator* of L (and note that $I \geq -1$).

Psychophysically, it is well known that, over substantial ranges of μ, the apparent motion of L does not depend on μ

Thus the contrast modulator, I of L emerges as a likely function to analyze for the motion information carried by L. Accordingly, we shall shift our focus from luminance to contrast and identify L with its contrast modulator, dropping reference to adaptation level.

Specifically, we shall call any function $I: Z^3 \to \mathbb{R}$ a *stimulus* iff $I[x, y, t] = 0$ for all but finitely many points $(x, y, t) \in Z^3$.

Strictly speaking, we should also require that I never drop below -1. This restriction, however, would lead to unnecessary complications in dealing with various sorts of combinations of stimuli. In all cases, the points that we wish to make tolerate rescaling of stimuli by arbitrary multiplicative constants to settle their minimal values to some perceptually appropriate level between -1 and 0. Accordingly, we drop the restriction that $I \geq -1$.

In general, we shall consider stimuli of two spatial dimensions and time. The reader may find it convenient to think of the first spatial dimension (which we shall always index by x) as horizontal, with indices increasing to the right, and the second spatial dimension (always indexed by y) as vertical, with indices increasing upward. The temporal dimension is always indexed by t.

Frames and Frame Blocks

For any stimulus I, we call the restriction of I to $Z^2 \times \{t\}$ the tth *frame* of I. In all the stimulus examples that we shall consider, frames clump into blocks: specifically, for each demonstration stimulus I defined in this paper, there are integers k and N such that all changes in luminance occur in frames kn, where $n = 0, 1, \ldots, N$, and otherwise luminance remains constant across frames. The group of identical frames between and including frames kn and $kn + k - 1$ we shall call the nth *frame block* of I.

Any stimulus I is nonzero at only a finite number of points in its countably infinite domain. Consequently, (i) the mean value of I is 0, and (ii) the power in I is finite.

From property (ii) we observe that I has a well-defined Fourier transform, which we denote by $\mathcal{I}$. Specifically,

$$\mathcal{I}(\omega, \theta, \tau) = \sum_{(x,y,t) \in Z^3} I[x, y, t] \exp(-j(\omega x + \theta y + \tau t)) \quad \text{(analysis)}.$$

We shall always use square brackets around the arguments of discrete functions and parentheses around the arguments of continuous functions. Although $\mathcal{I}$ is defined for all $(\omega, \theta, \tau) \in \mathbb{R}^3$, it is periodic over 2π in each variable. This fact is reflected in the inverse transform:

$$I[x, y, t] = \frac{1}{(2\pi)^3} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \mathcal{I}(\omega, \theta, \tau) \times \exp(j(\omega x + \theta y + \tau t)) d\omega d\theta d\tau \quad \text{(synthesis)}.$$

In the Fourier domain we shall consistently use ω to index frequencies relative to x, θ to index frequencies relative to y, and τ to index frequencies relative to t. This convention is exemplified by the definition of $\mathcal{I}$ above.

We distinguish the stimulus 0 by setting $0[x, y, t] = 0$ for all $x, y, t \in Z$. In parallel, we let $\mathcal{0}$ assign 0 to all $(\omega, \theta, \tau) \in \mathbb{R}^3$.

5. DRIFT-BALANCED RANDOM STIMULI

We begin by generalizing the notion of a stimulus to that of a random stimulus. Whereas a nonrandom stimulus assigns fixed values to Z^3, a random stimulus I assigns jointly distributed random variables that deviate from zero at only a finite number of points.

Various expectations associated with I are defined easily. We shall be particularly interested in the expected power of I at some point, (ω, θ, τ) in the frequency domain: $E[|\mathcal{I}(\omega, \theta, \tau)|^2]$. This reflects the expected power in I of a sinusoid C that modulates contrast at the rate of $\omega/2\pi$ cycles per column, $\theta/2\pi$ cycles per row, and $\tau/2\pi$ cycles per frame. The sinusoid with the same spatial frequency as C and moving at the same rate but in the opposite direction is obtained simply by reversing the direction of C's temporal contrast modulation: that is, by modulating contrast $-\tau/2\pi$ cycles per frame. When the expected power in I of any given drifting sinusoid is matched by the expected power of the sinusoid of the same spatial frequency drifting at the same rate in the opposite direction, we call I drift balanced.

Although the MFFC principle suggests that drift-balanced random stimuli should not display consistent apparent motion across independent realizations, we shall provide examples of drift-balanced random stimuli (in Section 6) that do in fact display strong, consistent motion across trials.

Beyond these basic developments, two propositions are proved in this section. In proposition 1 we demonstrate that any random stimulus separable in space and time (see definition 3 below) is drift balanced, and in proposition 2 we show that the (spatiotemporal) convolution of any two independent, drift-balanced random stimuli is drift balanced.

We now proceed more precisely as follows.

Definition 1: Random Stimulus

Call any family $I[x, y, t]$, $(x, y, t) \in Z^3$, of random variables jointly distributed with density f, a *random stimulus* when

(i) $I[x, y, t] = 0$ for all but a finite subset $\alpha \subset Z^3$ and
(ii) $E[I[x, y, t]^2]$ exists for all $(x, y, t) \in Z^3$.

Expectations Related to I

With k the cardinality of α, we set up a one-to-one correspondence between dimensions of $\mathbb{R}^k$ and points of α so that each coordinate of any vector $i \in \mathbb{R}^k$ corresponds to one of the points of α. We can now treat i as a stimulus (whose nonzero values are restricted to the points of α). In particular, letting $i_{(p,q,r)}$ denote the coordinate of i corresponding to a given $(p, q, r) \in \alpha$, we set

$$i[x, y, t] = \begin{cases} i_{(x,y,t)} & \text{if } (x, y, t) \in \alpha \\ 0 & \text{otherwise} \end{cases}$$

for any $(x, y, t) \in Z^3$. We can now conveniently formulate various expectations associated with I; in particular, we define the expectation of I by

$$E_I[x, y, t] = \int_{\mathbb{R}^k} i[x, y, t] f(i) di$$

for all $(x, y, t) \in Z^3$. (Note that E_I is a nonrandom stimulus.)

Consider the Fourier transform of E_I:

$$E_I(\omega,\theta,\tau) = \sum_{(x,y,t)\in Z^3} \int_{R^1} i[x,y,t]f(i)di \times \exp(-j(\omega x+\theta y+\tau t))$$

$$= \int_{R^1} \sum_{(x,y,t)\in Z^3} i[x,y,t]\exp(-j(\omega x+\theta y+\tau t))f(i)di$$

$$= \int_{R^1} \tilde{I}(\omega,\theta,\tau)f(i)di = E[\tilde{I}(\omega,\theta,\tau)].$$

This leads to the following observation.

Observation 1
The Fourier transform of the expectation of a random stimulus I is equal to the expectation of the Fourier transform of I.

Note especially, here, the implication that $E_I = 0$ iff $E[\tilde{I}] = \tilde{0}$.

We call any random stimulus I *invariant* iff there exists a stimulus S such that $I = S$ with probability 1.

Example 2: Randomly Contrast-Reversing, Rightward-Stepping Vertical Bar
Let the random stimulus I contain four frame blocks indexed 0, 1, 2, and 3, and let each frame block be composed of a horizontal sequence of four rectangles indexed 0, 1, 2, and 3 from left to right. Let ϕ_0, ϕ_1, ϕ_2, and ϕ_3 be pairwise independent random variables, each taking the value C or $-C$ with equal probability. Give rectangle i in frame block i the value assumed by ϕ_i, and give all other pixels the value 0.

The restriction of I to any one of its rows is characterized by Fig. 3, as a function of x along the horizontal axis and t along the vertical axis. As is clear, for any $(x, y, t) \in Z^3$,

$$E[I[x,y,t]] = 0;$$

that is, $E_I = 0$, from which we infer that $\tilde{E}_I = \tilde{0}$.

An interesting fact that may not be so obvious, however, (this follows from corollary 1 below) is that the expected power contributed to I by any given drifting sinusoidal grating is equal to the expected power contributed by the grating of the same spatial frequency drifting at the same rate in the opposite direction. This may seem surprising in light of the MFFC principle, since any realization of I is marked by a systematic, left-to-right perturbation across time, which (as one might expect) tends, under appropriate viewing conditions, to be perceived as motion from left to right. Indeed, as we shall see in Section 6, it is quite easy to construct random stimuli with this property that nonetheless display striking, reliable apparent motion in a fixed direction.

This fact motivates a notion central to this paper: that of a drift-balanced random stimulus (see definition 2 below). As the name suggests, a drift-balanced random stimulus is one for which the expected contribution of any given drifting sinusoidal grating is balanced by (equal to) the expected contribution of the corresponding grating drifting at the same rate in the opposite direction. Of course, just as a given random variable may have little or no probability of assuming a value equal to its expectation, a particular realization of a drift-balanced random stimulus, I, does not, in general, have perfectly balanced components. However, when gauged over a number of independent realizations, the mean contribution of a particular Fourier component of I tends to balance against the contribution of the corresponding, oppositely moving component.

	rectangle 0	1	2	3
frame block 3	0	0	0	ϕ_3
frame block 2	0	0	ϕ_2	0
frame block 1	0	ϕ_1	0	0
frame block 0	ϕ_0	0	0	0

Fig. 3. Rightward-stepping, randomly contrast-reversing vertical bar: a horizontal-temporal diagram of the random stimulus I, a vertical bar that appears with contrast C or $-C$ randomly assigned and steps its width rightward three times over a zero-contrast visual field, assuming contrast C or $-C$ with equal probability with each step. The expected power in I of any given drifting sinusoid is equal to the expected power of the sinusoid of the same spatial frequency drifting at the same rate but in the opposite direction.

Definition 2: Drift-Balanced Random Stimulus
Call any random stimulus I *drift balanced* iff, for any $\omega, \theta, \tau \in \mathbb{R}$,

$$E[|\tilde{I}(\omega,\theta,\tau)|^2] = E[|\tilde{I}(\omega,\theta,-\tau)|^2]. \quad (2)$$

[For a proof that the expectations in Eq. (2) exist, see Appendix A.] Notice that, because I is real valued, Eq. (2) is equivalent to

$$E[|\tilde{I}(\omega,\theta,\tau)|^2] = E[|\tilde{I}(-\omega,-\theta,\tau)|^2];$$

that is, I is drift balanced iff the expected power in I of any given drifting sinusoidal grating is equal to the expected power of the grating with the same spatial frequency drifting at the same rate but in the opposite direction.

As we shall see in Section 6, the following class of random stimuli is useful in constructing drift-balanced random stimuli that display consistent motion.

Definition 3: Space-Time-Separable Random Stimulus
Call any random stimulus I *space-time separable* iff, for any $(x, y, t) \in Z^3$,

$$I[x,y,t] = g[x,y]h[t],$$

for jointly distributed real random functions g and h.

Immediately we note a simple proposition.

Proposition 1
Any space-time-separable random stimulus is drift balanced.

Proof
Let I be a space-time-separable random stimulus, with

$$I[x,y,t] = g[x,y]h[t]$$

for all $(x, y, t) \in \mathbb{Z}^3$; then

$$|\overline{I(\omega, \theta, \tau)}|^2 = |\overline{g}(\omega, \theta)|^2 |\overline{h}(\tau)|^2.$$

Thus, since h is real valued,

$$|I(\omega, \theta, \tau)|^2 = |\overline{g}(\omega, \theta)|^2 |\overline{h}(-\tau)|^2 = |I(\omega, \theta, -\tau)|^2.$$

Taking expectations of both sides yields Eq. (2). ∎

It would be surprising for any space–time-separable random stimulus I to exhibit strong, consistent motion in a fixed direction, since the only sort of temporal contrast change induced by I is a spatially global modulation.

However, as we have hinted in example 1, there do exist drift-balanced random stimuli that exhibit decisive motion in a fixed direction not only on the average across a number of trials but on virtually each display. In Section 6 we shall provide some general results that are useful for constructing a broad range of drift-balanced random stimuli that show strong motion. However, we shall show first that the spatiotemporal convolution of independent drift-balanced random stimuli is drift balanced and briefly mention some of the ramifications of this fact.

Proposition 2
The (spatiotemporal) convolution of independent, drift-balanced random stimuli is drift balanced.

Proof
Let I and J be independent drift-balanced random stimuli. For any random stimuli we have

$$|\overline{I * J}|^2 = |\overline{I}|^2 |\overline{J}|^2.$$

The independence of I and J implies that

$$E[|\overline{I}|^2 |\overline{J}|^2] = E[|\overline{I}|^2] E[|\overline{J}|^2].$$

Thus, since I and J are drift balanced, we find that, for any ω, $\theta, \tau \in \mathbb{R}$,

$$\begin{aligned} E[|\overline{I * J}(\omega, \theta, \tau)|^2] &= E[|\overline{I}(\omega, \theta, \tau)|^2] E[|\overline{J}(\omega, \theta, \tau)|^2] \\ &= E[|\overline{I}(\omega, \theta, -\tau)|^2] E[|\overline{J}(\omega, \theta, -\tau)|^2] \\ &= E[|\overline{I * J}(\omega, \theta, -\tau)|^2]. \end{aligned}$$ ∎

Most computational models of the sensory transformations mediating human perception routinely apply a spatiotemporal, linear, shift-invariant filter to the input stimulus. The impulse response (i.e., convolution kernel) of any such filter can, of course, be regarded as an invariant stimulus. Typically the filters applied are drift balanced.[1–4,17] Obviously, filters that depend on only spatial characteristics of the stimulus being processed are drift balanced (for instance, all manner of oriented, band-tuned, spatial edge detectors). Similarly, filters (such as flicker detectors) that depend on only temporal stimulus characteristics are drift balanced. More generally, all space–time-separable filters are drift balanced (proposition 1). Thus, given a drift-balanced random input stimulus, the output of many of the filters that are commonly thought to function in the early stages of human visual processing is also drift balanced.

6. CONSISTENT APPARENT MOTION FROM DRIFT-BALANCED STIMULI

We begin this section by noting some general results concerning linear combinations of random stimuli, leading up to proposition 3 below, in which we show that any linear combination of pairwise independent, drift-balanced random stimuli, all of which have expectation 0, is drift balanced. (Actually, this is an implication of proposition 3, which is slightly more general.) From this finding follow corollaries 1 and C1 (C1 in Appendix C), each of which gives rise to specific examples of drift-balanced random stimuli that elicit consistent apparent motion. Several of these examples are detailed in this section. Experimental findings with regard to these example random stimuli are reported.

One may wonder whether linear combinations of independent drift-balanced random stimuli are drift balanced. That this is not the case is evident from the fact that any invariant stimulus whatsoever can be expressed as a linear combination of shifted impulses, which are, of course, jointly independent and individually drift balanced.

Although linear combinations of arbitrary, pairwise independent, drift-balanced random stimuli are not generally drift balanced, if we impose an additional constraint on the random stimuli to be summed we can ensure that the resultant linear combination is indeed drift balanced.

The following lemma bears on this issue.

Lemma 1
Let S be a random stimulus equal to the sum of a set Ω of pairwise independent random stimuli; then

$$E[|S|^2] = |E_s|^2 + \sum_{I \in \Omega} E[|N_I|^2],$$

where $N_I = I - E_I$ for each $I \in \Omega$.

Proof
See Appendix B.

Immediately we note a useful result concerning linear combinations of drift-balanced random stimuli:

Proposition 3
Let $\Omega = \Theta \cup \{I\}$ be a set of pairwise independent, drift-balanced random stimuli, such that I is invariant and each member of Θ has an expectation of 0. Then any linear combination, S, of the elements of Ω is drift balanced.

Proof
A drift-balanced random stimulus rescaled by a constant is drift balanced. Thus we assume with no loss of generality that S is just a sum of pairwise independent drift-balanced random stimuli.

Note that (i) $I = E_I$ (hence $N_I = I - E_I = 0$) and (ii) for all $J \in \Theta$, $N_J = J - E_J = J$. Thus from lemma 1 we observe for any $\omega, \theta, \tau \in \mathbb{R}$

$$E[|S(\omega,\theta,\tau)|^2] = |E_S(\omega,\theta,\tau)|^2 + \sum_{J\in\mathbb{Z}} E[|S_J(\omega,\theta,\tau)|^2]$$

$$= |J(\omega,\theta,\tau)|^2 + \sum_{J\in\Theta} E[|J(\omega,\theta,\tau)|^2]$$

$$= |J(\omega,\theta,-\tau)|^2 + \sum_{J\in\Theta} E[|J(\omega,\theta,-\tau)|^2]$$

$$= E[|S(\omega,\theta,-\tau)|^2]. \quad \blacksquare$$

Note, in particular, that this result holds for $I = \Theta$.

As is reasonably clear from proposition 3 (since space-time-separable random stimuli are drift balanced), any sum of pairwise independent, space-time-separable random stimuli, all with an expectation of Θ, is drift balanced. In corollary 1 this principle is applied to generate a class of drift-balanced random stimuli, certain instances of which exhibit strong, consistent apparent motion in a fixed direction.

Corollary 1

For $M \in Z^+$, let $\phi_0, \phi_1, \ldots, \phi_{M-1}$ be pairwise independent random variables, each with expectation 0; and, for $m = 0, 1, \ldots, M-1$, let $f_m: Z^2 \to \mathbb{R}$ and $g_m: Z \to \mathbb{R}$, and let the product $f_m g_m$ be 0 at all but finitely many points of Z^2; then the random stimulus I defined by setting

$$I[x,y,t] = \sum_{m=0}^{M-1} \phi_m f_m[x,y] g_m[t], \qquad (3)$$

is drift balanced.

The proof is obvious from propositions 1 and 3.

A simple yet compelling counterexample to the MFFC principle may now be constructed as follows.

Demonstration 1: A Randomly Contrast-Reversing, Rightward-Stepping Rectangle

For some $M \in Z^+$, let the random stimulus I be composed of M frame blocks indexed $0, 1, \ldots, M-1$, and let each frame block be composed of a horizontal sequence of M rectangles indexed $0, 1, \ldots, M-1$ from left to right (see example 2 and Fig. 3). Let $\phi_0, \phi_1, \ldots, \phi_{M-1}$ be pairwise independent random variables, each taking the value C or $-C$ with equal probability. Give rectangle i in frame block i the value assumed by ϕ_i, and give all other pixels the value 0. We can now define I by Eq. (3) by letting $f_m[x,y]$ take the value 1 in the mth rectangle and 0 everywhere else and letting $g_m[t]$ take the value 1 in the mth frame block and 0 everywhere else.

The apparent motion of this stimulus is quite easy to imagine: throughout frame block 0, rectangle 0 is present on the left-hand side of the stimulus field, it is assigned contrast of C or $-C$ with equal probability. In frame block 1, rectangle 0 turns off (goes to contrast 0), and rectangle 1, abutting rectangle 0 on the right, turns on, again with contrast C or $-C$ assigned with equal probability, independent of the contrast of the first rectangle. In each successive frame block, one rectangle turns off, and a new rectangle turns on directly to the right of its predecessor, with contrast either C or $-C$, independent of any other rectangle.

Figure 4a displays a realization of one version of the random stimulus I defined in demonstration 1 with $M = 8$. This random stimulus and others that we shall discuss were tested experimentally on two subjects. Before discussing responses to I in particular, we describe the experimental arrangements for these observations.

General Method

We describe here the procedure for demonstrations 1 (stimulus *I*), 2 (*K*), 3 (*J*), 4 (*H*), and 5 (*G*). All stimulus presentations were made on a Conrac 7211 RGB monitor driven by an Adage graphics display processor. The display area was 28 cm × 32 cm, and displayed intensities were greenish white. The spatial resolution was 512 × 512 pixels, the temporal resolution was 60 frames/sec, and the intensity resolution was 256 gray levels.

Two subjects were involved in each of the studies: CC (the experimenter) and DY (a naive subject). For each demonstration, each subject viewed 30 independent realizations of the random stimulus. On each presentation, the non-Fourier motion of the stimulus (*I*, *K*, *J*, *H*, or *G*) was left to right or right to left with equal probability. For instance, *I*'s randomly contrast-reversing rectangle stepped left to right or right to left with equal probability.

Subjects adapted before each session to a uniform screen of luminance 60 cd/m^2; other luminances were linearized carefully relative to the mean. All stimuli were viewed foveally and binocularly, from a distance of 2 m. On each trial, a central cue spot (0.5 deg × 0.5 deg) of low positive contrast came on 2 sec before the onset of the stimulus and disappeared 1 sec before the onset. Subjects were instructed to maintain their gazes throughout the trial on the cue spot point and were required to indicate the predominant direction of apparent motion (left or right) by entering either an L or an R on a terminal keyboard.

Method for Demonstration 1

In the version of I viewed by our subjects, frame blocks lasted 1/60 sec; spatial rectangles measured approximately 2 deg (horizontal) × 2 deg (vertical) and $C = 0.25$. The contrast of 0.25 was chosen because it produced easily visible motion and yet was small enough that psychophysical, as well as physical, equivalence of positive and negative increments was likely to hold.

Results

Subject CC (DY) reported apparent motion in the step direction on 30 (29) of 30 trials.

Discussion

The essential trick of the rightward stepping bar was to modulate the contrast (that is, the absolute deviation from zero) of a field of static, spatially independent, zero-mean noise as a function of space and time. This notion of spatio temporal modulation of contrast needs some explanation Let J be a random stimulus with expectation 0, let W be a nonnegative function of Z^3 (space and time), and consider $I = WJ$. In general, J's distance from 0, be it positive or negative, is magnified (or damped) by W's value at each point in space and time. Thus I is obtained by letting W modulate the (absolute) contrast of J.

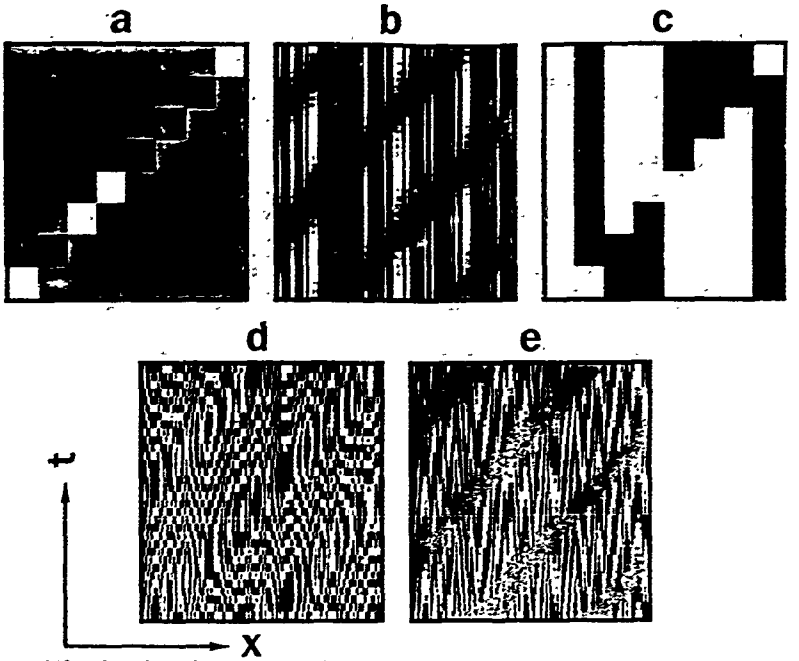


Fig. 4. a, Rightward-stepping, randomly contrast-reversing vertical bar: a horizontal-temporal cross section of a realization of the random stimulus I (see demonstration 1). I is the sum of pairwise independent space-time-separable random stimuli, each of which has an expectation of 0; consequently I is drift balanced (by corollary 1). b, Modulation of the contrast of a static noise field by a drifting sinusoidal grating: a horizontal-temporal cross section of a realization of the random stimulus K (demonstration 2). That K is drift balanced follows from corollary 1. c, Traveling contrast reversal of a noise field: a horizontal-temporal cross section of a realization of the random stimulus J (demonstration 3). J is the sum of pairwise independent space-time-separable random stimuli, each of which has an expectation of 0 and is thus drift balanced (by corollary 1) Note that, in contrast to $|I|$ (for I of Fig 4a), $|J|$ is devoid of motion information. d, Modulation of the flicker frequency of a flickering noise field by a drifting grating: a horizontal-temporal cross section of a realization of the random stimulus H (demonstration 4). That H is drift balanced is a consequence of corollary C1 (in Appendix C). The motion of H is derived from spatiotemporal modulation of the frequency of sinusoidal flicker, where the phase of the flicker is random over space. e, Modulation of the contrast of a flickering noise field by a drifting sinusoidal grating: a horizontal-temporal cross section of a realization of the random stimulus G (demonstration 5) G is drift balanced (by corollary C1). The motion of G is derived from spatiotemporal modulation of the amplitude of sinusoidal flicker, where the flicker phase is random over space.

To see how this notion applies to I of demonstration 1, note that we can look at I as the result of multiplying a field J of random black or white rectangles persisting through M chunks of time by a function W, which (for $m = 0, 1, \ldots M-1$) is 1 in the mth frame block for the points in the mth rectangle from the left and 0 everywhere else.

Elaborations of this basic contrast-modulation scheme are easy to construct. Consider, for instance, demonstration 2.

Demonstration 2: Contrast Modulation of a Static Noise Field by a Drifting Sinusoid

We compose the random stimulus K of N frame blocks, each containing a horizontal row of rectangles, indexed $0, 1, \ldots, M-1$ from left to right. For $m = 0, \ldots, M-1$, let $f_m[x, y]$ take the value 1 in the mth rectangle and 0 elsewhere, and let $g_m[t]$ vary as a sinusoidal function of m and the frame block. Specifically, for each frame t in the nth frame block, let

$$g_m[t] = \frac{\cos[2\pi(\alpha m/M - \beta n/N)] + 1}{2}$$

for some spatial and temporal frequencies α and β. Let $\phi_0, \phi_1, \ldots, \phi_{M-1}$ be pairwise independent random variables taking the values C and $-C$ with equal probability, for some contrast C, and define K by Eq. (3).

Whereas I of demonstration 1 merely picks out successive rectangles of spatial noise (independently assigned contrast C or $-C$) in successive time intervals, K is marked by high-power crests (α per frame block) separated by zero-power (gray) troughs sweeping at a constant rate from left to right over the row of rectangles, each of random contrast C or $-C$.

Figure 4b shows a realization of K, with $M = 128$, $N = 32$, and $\alpha = \beta = 2$.

Method

In the version of K viewed by our subjects, frame blocks lasted 1/60 sec, rectangles measured approximately (1/8 deg horizontal) by (2 deg vertical), and contrast $C = 0.25$.

Results

The cosine grating modulating the contrast of K was rightward or leftward drifting with equal probability. Subject CC (DY) reported apparent motion in the direction of drift in 30 (26) of 30 trials.

It might be that humans extract the motion from stimuli such as I (Fig. 4a) and K (Fig. 4b) simply by performing a Fourier power analysis on a rectified version of the stimulus. For instance, if subjects were able either (i) to disregard (set to 0) all negative contrast values or (ii) to map all contrasts onto their absolute values, then it is clear that a Fourier power analysis of the resultant rectified signal would correspond quite well to perceived motion. This explanation does not account for responses to stimuli of the type considered in demonstration 3.

Demonstration 3: Traveling Contrast Reversal of a Random Bar Pattern

Let $M \in Z^+$. We construct the random stimulus J of $M + 1$ frame blocks indexed $0, 1, \ldots, M$, each of which contains M rectangles indexed $0, 1, \ldots, M - 1$ from left to right. Let $I_m[x, y]$ take the value 1 in the mth rectangle and zero elsewhere; let $g_m[t]$ be 1 in frame blocks 0 through m, -1 in frame blocks $m + 1$ through M, and 0 everywhere else. Let the random variables $\phi_0, \phi_1, \ldots, \phi_{M-1}$ be pairwise independent, each taking a contrast value of C or $-C$ with equal probability, and use Eq. (3) to define J.

In frame block 0 of J, all M rectangles turn on, some with contrast C and others with contrast $-C$. In successive frame blocks $m = 1, 2, \ldots, M$, exactly one of the rectangles changes contrast: the $(m-1)$th switches to C if its previous contrast was $-C$; otherwise it flips from C to $-C$. In frame block 1, the leftmost (0th) rectangle flips contrast; in frame block 2, rectangle 1 flips, and in successive frame blocks, successive rectangles flip contrast from left to right, until the $(M-1)$th rectangle flips in frame block M, after which all the rectangles turn off.

Method

The version of J viewed by subjects CC and DY contained nine frame blocks, each of which lasted 1/60 sec and contained eight spatial rectangles, each measuring approximately 2 deg × 2 deg; $C = 0.25$.

Results

CC (DY) reported apparent motion in the direction traveled by the contrast flip in 30 (25) of 30 trials.

The next two stimuli (G of demonstration 4 and H of demonstration 5) are both drift balanced. The proof of this fact depends on a corollary to proposition 3 that is otherwise unimportant. We relegate this corollary to Appendix C and show there how it can be applied to construct each of G and H.

Demonstration 4: Modulating the Flicker Frequency of Spatial Noise with a Drifting Sinusoid

We shall construct the random stimulus H of N frame blocks indexed $0, 1, \ldots, N-1$, each composed of M rectangles indexed $0, 1, \ldots, M-1$ from left to right. Let $\rho_0, \rho_1, \ldots, \rho_{M-1}$ be pairwise independent random variables, each uniformly distributed on $[-\pi, \pi)$. Let C be a contrast value. For all $(x, y, t) \in Z^3$, set

$$H[x, y, t] = C \cos\left(4\pi\left(1 + \cos\left(2\pi\left(\frac{m}{M} - \frac{n}{N}\right)\right)\right) + \rho_m\right)$$

for m indexing the rectangle containing (x, y) and n indexing the frame block containing t. The demonstration that H is drift balanced is given in Appendix C.

A realization of H, with $N = 32$ and $M = 128$, is shown in Fig. 4d. In frame block 0, the rectangles are assigned random contrasts between C and $-C$ (as a consequence of their independent, random phases). Thereafter, for $m = 0, 1, \ldots, M-1$, the contrast of the mth rectangle is modulated by a cosine whose phase is itself a sinusoidal function of the rectangle and the frame block. Since, however, a sinusoid's frequency is the derivative of its phase (and since the derivative of a sinusoid is a sinusoid of the same frequency), we observe that H modulates, with a drifting sinusoid, the frequency of (spatially random-phased) sinusoidal flicker.

In demonstration 4 the contrast oscillation rate of each rectangle speeds up and slows down sinusoidally throughout the presentation. Regions of equal oscillation rate (crests of rapid sinusoidal flicker separated by troughs of slow modulation) sweep at a constant rate from left to right across the viewing field.

Method

The conditions under which H was presented to subjects CC and DY were the same as those governing the display of K (of demonstration 2). Each frame block lasted 1/60 sec, each spatial rectangle measured 2 deg (vertical) × 1/8 deg (horizontal), and the contrast $C = 0.25$.

Results

Interestingly, despite the striking diagonal contours marking the (x, y) pattern of Fig. 4d, both subjects reported that the motion of H was generally more ambiguous than those of the other stimuli. CC (DY) reported apparent motion in the drift direction of the sinusoid modulating frequency of contrast oscillation on 28 (23) of 30 trials.

Demonstration 5: Modulating the Contrast of Flickering Noise with a Drifting Sinusoid

The random stimulus G is made up of N frame blocks indexed $0, 1, \ldots, N-1$, each containing M rectangles indexed $0, 1, \ldots, M-1$ from left to right. Let $\rho_0, \rho_1, \ldots, \rho_{M-1}$ be pairwise independent random variables, each uniformly distributed on $[-\pi, \pi)$. Let C be some contrast value, then, for any $(x, y, t) \in Z^3$, set

$$G[x, y, t] = \frac{C}{2}\left(\cos\left(2\pi\left(\alpha\frac{m}{M} - \beta\frac{n}{N}\right)\right) + 1\right) \times \cos\left(2\pi\gamma\frac{n}{N} + \rho_m\right),$$

where m indexes the rectangle containing (x, y) and n index-

es the frame block containing t. The proof that G is drift balanced is given in Appendix C.

A realization of G with $M = 128$, $N = 32$, $\alpha = \beta = 2$, and $\gamma = 3$ is shown in Fig. 4e. As does K of demonstration 2, G generates its apparent motion by modulating contrast as a drifting sinusoidal function of the rectangle and the frame block. However, whereas the background whose contrast is being modulated in K is a static row of rectangles randomly painted C or $-C$, the background whose power is modulated in G is a row of rectangles sinusoidally flickering between C and $-C$; each rectangle m has a randomly assigned phase (ρ_m) and is flickering at the rate of 3/32 cycles/frame block (as a consequence of the term $2\pi\ 3n/32$).

The contrast of G's flickering rectangle row is modulated by the factor

$$\cos\left(2\pi\left(\frac{2m}{128} - \frac{2n}{32}\right)\right) + 1,$$

which sweeps peaks (two per frame) of high-contrast flicker separated by troughs of mean gray across the viewing field from left to right.

Method

The conditions governing the display of G were the same as those for K (and H): Frame blocks lasted 1/60 sec, spatial rectangles measured 2 deg (vertical) × 1/8 deg (horizontal), and $C = 0.25$.

Results

CC (DY) registered apparent motion in the drift direction of the sinusoid modulating noise contrast in G on 30 (26) of 30 trials.

Conclusions

In this section we have demonstrated five drift-balanced random stimuli whose apparent motion is perceived in one consistent direction in more than 90% of trials by two observers. Indeed, many other observers have viewed these stimuli, and no one has yet failed to perceive their consistent motion. As is discussed in Section 8 below, these stimuli are microbalanced in addition to being drift balanced; that is, they remain drift balanced after windowing by arbitrary space–time-separable functions. We conclude that there is a large class of random stimuli whose apparent motion contradicts the MFFC principle of motion perception.

There are many kinds of drift-balanced and microbalanced random stimuli that were not represented among the demonstrations described here. In this paper we have restricted ourselves to stimuli that assign constant values in the vertical dimension of space. Dropping this constraint opens the door to a broad range of other drift-balanced and microbalanced random stimuli. In particular, a large class of displays that yield apparent motion is generated by defining two spatiotemporal texture fields, A and B, at each point $(x, y, t) \in Z^3$ and moving a boundary that admits light only from field A on one side and only from B on the other. Many instances of this kind of apparent motion, including those proposed by Victor,[18] can easily be shown to be microbalanced.[19]

7. REICHARDT-DETECTOR CHARACTERIZATION OF DRIFT-BALANCED RANDOM STIMULI

A point-delay Reichardt detector is a simple device that was proposed originally by Reichardt[20] to explain the vision of beetles. Its basic principle, the autocorrelation of inputs at nearby visual locations, underlies most of the currently predominant models of human motion perception. We define the Reichardt detector in terms of two subunits, designated for convenience as the left and right half-detectors. Both half-detectors are defined with respect to the same two (spatial) locations (x, y) and (p, q) in Z^2 and for some fixed nonnegative number δ_t of frames. These oppositely oriented detectors are pitted additively against each other. A left half-detector r_{left} [implicitly indexed by (x, y), (p, q), and δ_t] computes the covariance over time of the contrast at point (x, y) at time t with the contrast at point (p, q) at time $t - \delta_t$ throughout the display of an arbitrary stimulus I. For r_{right}, t and $t - \delta_t$ are reversed. The computation performed by r is given by

$$r(I) = r_{left}(I) - r_{right}(I) = \sum_{t \in Z} I[x, y, t] I[p, q, t - \delta_t] - \sum_{t \in Z} I[x, y, t - \delta_t] I[p, q, t].$$

When $r(I) < 0$, it indicates motion from (x, y) to (p, q).

Figure 5 illustrates a block-diagram representation of the Reichardt half-detectors and the Reichardt full detector. The box containing (x, y) [respectively, (p, q)] is a contrast gauge, inputting the contrast at point (x, y) [(p, q)] for each successive frame t. Each of the boxes containing δ_t is a delay filter. At frame t, each delay box outputs the value entered into it at frame $t - \delta_t$. Each of the boxes marked with an × outputs the product of its two inputs at any frame t. Each of the boxes marked with a $\sum$ accumulates the output from the multipliers over all the frames. Finally, the box marked with a − outputs the difference of its inputs at any frame t.

To see how the detector shown in Fig. 5c works, consider a point of light moving across a dark visual field so as to cross first (x, y) and then (p, q). If the spot is moving at the proper rate [so that it starts crossing (p, q) after precisely δ_t frames], then the output from the right-hand multiplier will be high as the dot passes over (p, q). In contrast, the output from the left-hand multiplier will be low throughout the presentation of the moving dot, since, at any frame, at least one of its input channels is contributing a value near zero. Thus the output of the detector is negative. On the other hand, if the dot passes first over (p, q) and then over (x, y), the detector's response is positive. In this simple case, the sign of the detector's output does a good job of signaling the direction of the dot's motion.

However, the point-delay Reichardt detector is highly vulnerable to aliasing. Imagine a train of evenly spaced dots passing at some speed s first over (x, y) and then over (p, q). For any s, it is easy to adjust the spacing between dots so that the output of the Reichardt detector of Fig. 5c signals rightward motion, leftward motion, or no motion at all.

Despite the shortcomings of the simple Reichardt detector, there is something appealing about its fundamental au-

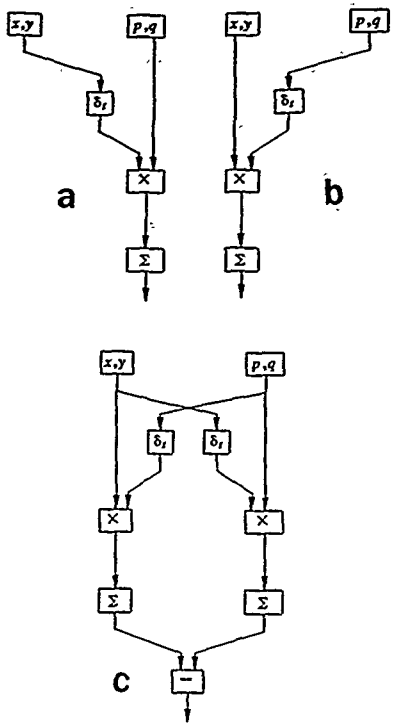


Fig. 5. Point-delay Reichardt detector and its component half-detectors. a, The right half-detector computes the covariance of the contrast fluctuations of the input stimulus at point (p, q) with the fluctuations δ_t frames earlier at point (x, y): (x, y) and (p, q) register signal contrast frame by frame. The contrast of the current frame at pixel (p, q) is multiplied by the contrast at pixel (x, y) δ_t frames in the past. (The box labeled δ_t outputs the input it received δ_t frames ago.) The output from the multiplier is accumulated over all the frames of the display. b, In a similar fashion, the left half-detector computes the covariance of the contrast fluctuations of the input stimulus at point (x, y) with the fluctuations δ_t frames earlier at point (p, q). c, The full point-delay Reichardt detector outputs the difference between the left and right half-detectors. A positive response thus signals leftward motion; a negative response signals rightward motion.

tocorrelation principle. Various elaborations of Reichardt models were developed and studied in detail by van Santen and Sperling,[1,2,21] who proved that the apparently different models of Adelson and Bergen[3] and Watson and Ahumada[4] were essentially special types of elaborated Reichardt detectors (ERD's). All these models retain the basic delay-and-compare structure of the simple detector diagrammed in Fig. 5c. However, this simple detector is generalized in the following ways: (i) the point detectors at (x, y) and (p, q) are replaced by spatial receptive fields (that is, each receptive field applies an array of weights to the stimulus impinging upon its region of the retina, and it outputs the sum of the weighted contrast values), (ii) the temporal point delays before the multipliers are replaced by temporal filters, and (iii) the temporal accumulators after the multipliers are replaced by temporal filters. Van Santen and Sperling[2] showed that further additions (e.g., more temporal filters added here and there) do not augment the capabilities of this ERD.

It was widely assumed that, ideally, a good motion detector should behave as a frequency-domain power analyzer.[1-6,21-23] (This is the assumption called into question by the demonstration of good apparent motion in drift-balanced stimuli.) The simple point-delay Reichardt detector falls short of this ideal: it is not a good Fourier analyzer. The various elaborations of Reichardt detectors can be viewed as attempts to improve their performance as frequency-domain power analyzers.

There is another way to use the Reichardt mechanism as the basis of a motion-perception model. Indeed, as we shall observe, it is possible to build a perfect Fourier power analyzer by using only the simplest point-delay half-detectors.

Our main purpose in this section, however, is to provide an alternative characterization of the class of drift-balanced random stimuli, in terms of the expected responses of point-delay Reichardt detectors to members of this class. We prove the following proposition: For any integers δ_x, δ_y, and δ_t, form the class $C_{\delta_x,\delta_y,\delta_t}$ of all point-delay Reichardt detectors conforming to Fig. 5c [with (x, y) and (p, q) ranging throughout Z^2] such that $(x, y) - (p, q) = (\delta_x, \delta_y)$, and call $C_{\delta_x,\delta_y,\delta_t}$ trivial if either $(\delta_x, \delta_y) = (0, 0)$ or $\delta_t = 0$; that is, $C_{\delta_x,\delta_y,\delta_t}$ is trivial if its member detectors fail to separate, either in space or time, the points whose contrast they compare. I is then drift balanced iff the expected pooled response of every nontrivial class of point-delay Reichardt detectors is 0. We now proceed more formally.

Definition 4: Autocorrelation
Let I be a random stimulus. Then for any $\delta = (\delta_x, \delta_y, \delta_t) \in Z^3$, define the *autocorrelation*, H_I, by

$$H_I[\delta_x, \delta_y, \delta_t] = \sum I[x, y, t]I[p, q, r],$$

where the sum is taken over all pairs $(x, y, t), (p, q, r) \in Z^3$ for which $(x, y, t) - (p, q, r) = (\delta_x, \delta_y, \delta_t)$. Define the *full-detector pooler*, R_I, by setting

$$R_I[\delta_x, \delta_y, \delta_t] = H_I[\delta_x, \delta_y, \delta_t] - H_I[-\delta_x, -\delta_y, \delta_t].$$

We use H_I to denote the autocorrelation of I because, for any $(\delta_x, \delta_y, \delta_t)$, $H_I[\delta_x, \delta_y, \delta_t]$ collects the sum of the responses to I of all the half-detectors conforming to Fig. 5b, with δ_t delay filters, such that $(x, y) - (p, q) = (\delta_x, \delta_y)$. The half-detectors corresponding to Fig. 5a are pooled by $H_I[-\delta_x, -\delta_y, \delta_t]$. Thus $R_I[\delta_x, \delta_y, \delta_t]$ pools the output of all full Reichardt detectors corresponding to Fig. 5c, with $(x, y) - (p, q) = (\delta_x, \delta_y)$ (and δ_t delay filters).

Observation 2
For any random (or nonrandom) stimulus I and any $\delta = (\delta_x, \delta_y, \delta_t) \in Z^3$,

$$\mathbf{H}_I[\delta] = \mathbf{H}_I[-\delta].$$

The proof is trivial.

In order to reclaim Fourier motion information from the half-detector output, note first that, for any random stimulus I,

$$|I(\omega,\theta,\tau)|^2 = \sum I[x,y,t]I[p,q,r] \times \exp(j(\omega(x-p)+\theta(y-q)+\tau(t-r))), \quad (4)$$

where the sum is taken over all $(x, y, t), (p, q, r) \in Z^3$. We can now collect terms of the sum in Eq. (4) that have identical exponential factors to obtain

$$|I(\omega,\theta,\tau)|^2 = \sum \mathbf{H}_I[\delta_x,\delta_y,\delta_t]\exp(j(\omega\delta_x+\theta\delta_y+\tau\delta_t)), \quad (5)$$

where the sum is over all $(\delta_x, \delta_y, \delta_t) \in Z^3$.

Equation (5) shows that point-delay half-detectors, by themselves, contain all the information about the distribution of I's power in the Fourier domain (because $\mathbf{H}_I$ depends on only the output of half-detectors to I).

The next definition is useful for proving the main result of this section.

Definition 5: Power Difference between Oppositely Drifting Fourier Components

For any random stimulus I and any $\omega, \theta, \tau \in \mathbb{R}$, set

$$\Delta_I(\omega,\theta,\tau) = |I(\omega,\theta,\tau)|^2 - |I(\omega,\theta,-\tau)|^2.$$

Note that any random stimulus I is drift balanced iff $E[\Delta_I(\omega,\theta,\tau)] = 0$ for all $\omega, \theta, \tau \in [0, 2\pi)$. Some facts about Δ_I are worth noting First,

$$\begin{aligned}\Delta_I(\omega,\theta,\tau) &= \sum(\mathbf{H}_I[\delta_x,\delta_y,\delta_t] - \mathbf{H}_I[\delta_x,\delta_y,-\delta_t]) \times \exp(j(\omega\delta_x+\theta\delta_y+\tau\delta_t)) \\ &= \sum(\mathbf{H}_I[\delta_x,\delta_y,\delta_t] - \mathbf{H}_I[-\delta_x,-\delta_y,\delta_t]) \times \exp(j(\omega\delta_x+\theta\delta_y+\tau\delta_t)) \\ &= \sum \mathbf{R}_I[\delta_x,\delta_y,\delta_t]\exp(j(\omega\delta_x+\theta\delta_y+\tau\delta_t)),\end{aligned}$$

where each sum is over all $(\delta_x, \delta_y, \delta_t) \in Z^3$. The first identity depends on the fact that

$$\begin{aligned}|I(\omega,\theta,-\tau)|^2 &= \sum \mathbf{H}_I[\delta_x,\delta_y,\delta_t]\exp(j(\omega\delta_x+\theta\delta_y-\tau\delta_t)) \\ &= \sum \mathbf{H}_I[\delta_x,\delta_y,-\delta_t]\exp(j(\omega\delta_x+\theta\delta_y+\tau\delta_t)).\end{aligned}$$

The second identity follows from observation 2.

Next note that any term

$$(\mathbf{H}_I[\delta_x,\delta_y,\delta_t] - \mathbf{H}_I[\delta_x,\delta_y,-\delta_t])\exp(j(\omega\delta_x+\theta\delta_y+\tau\delta_t))$$

in the sum yielding $\Delta_I(\omega, \theta, \tau)$ is obviously 0 if $\delta_t = 0$. On the other hand, this term is equal (by observation 2) to

$$(\mathbf{H}_I[\delta_x,\delta_y,\delta_t] - \mathbf{H}_I[-\delta_x,-\delta_y,\delta_t])\exp(j(\omega\delta_x+\theta\delta_y+\tau\delta_t)),$$

which is evidently 0 if $\delta_x = \delta_y = 0$. This goes to show that for any $\delta_x, \delta_y, \delta_t \in Z$, any class of Reichardt half-detectors, each of whose members has (i) no separation between spatial receptors or (ii) a delay factor of 0, does not influence $\Delta_I(\omega, \theta, \tau)$.

The following lemma summarizes these observations

Lemma 2

For any random stimulus I, any $\omega, \theta, \tau \in \mathbb{R}$,

$$\Delta_I(\omega,\theta,\tau) = \sum \mathbf{R}_I[\delta_x,\delta_y,\delta_t]\exp(j(\omega\delta_x+\theta\delta_y+\tau\delta_t)), \quad (6)$$

where the sum is taken over all integers δ_x, δ_y, and δ_t such that $\delta_t \neq 0$ and either $\delta_x \neq 0$ or $\delta_y \neq 0$.

Obviously, if

$$E[\mathbf{R}_I[\delta_x,\delta_y,\delta_t]] = 0$$

for all δ_x, δ_y, and δ_t indexing the sum in Eq. (6), then $\Delta_I(\omega, \theta, \tau) = 0$. This proves half of the following proposition.

Proposition 4

A random stimulus is drift balanced iff the expected pooled output from every nontrivial class of Reichardt detectors is 0; that is, any random stimulus I is drift-balanced iff

$$E[\mathbf{R}_I[\delta_x,\delta_y,\delta_t]] = 0 \quad (7)$$

for all integers δ_x, δ_y, and δ_t such that $\delta_t \neq 0$ and $(\delta_x, \delta_y) \neq (0, 0)$.

Proof

We have already observed that Eq. (7) implies that I is drift balanced. It remains to be proved that Eq. (7) holds whenever I is drift balanced. Accordingly, let Q be the set of all $(\delta_x, \delta_y, \delta_t)$ for which $\delta_t \neq 0$ and $(\delta_x, \delta_y) \neq (0, 0)$, and suppose that, for any $\omega, \theta, \tau \in [0, 2\pi)$,

$$E[\Delta_I(\omega,\theta,\tau)] = 0.$$

When we take expectations of both sides of Eq. (6), and multiply each side of the resulting identity by its conjugate, we obtain

$$E^2[\Delta_I(\omega,\theta,\tau)] = \sum E[\mathbf{R}_I[\delta_x,\delta_y,\delta_t]]E[\mathbf{R}_I[\delta_p,\delta_q,\delta_r]] \times \exp(j(\omega(\delta_x-\delta_p)+\theta(\delta_y-\delta_q)+\tau(\delta_t-\delta_r))), \quad (8)$$

where the sum is over all $(\delta_x, \delta_y, \delta_t), (\delta_p, \delta_q, \delta_r) \in Q$. However, recalling that

$$\begin{aligned}\int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi} &\exp(j(\omega(\delta_x-\delta_p)+\theta(\delta_y-\delta_q)+\tau(\delta_t-\delta_r)))\mathrm{d}\omega\mathrm{d}\theta\mathrm{d}\tau \\ &= \int_0^{2\pi}\exp(j\omega(\delta_x-\delta_p))\mathrm{d}\omega\int_0^{2\pi}\exp(j\theta(\delta_y-\delta_q))\mathrm{d}\theta \times \int_0^{2\pi}\exp(j\tau(\delta_t-\delta_r))\mathrm{d}\tau \\ &= \begin{cases}(2\pi)^3 & \text{if } \delta_x=\delta_p,\quad \delta_y=\delta_q,\quad \delta_t=\delta_r \\ 0 & \text{otherwise}\end{cases}\end{aligned}$$

we find that when we integrate both sides of Eq. (8) over the interval $[0, 2\pi)^3$ and divide through by $(2\pi)^3$, we obtain

$$\sum E^2[\mathbf{R}_I[\delta_x,\delta_y,\delta_t]] = \frac{1}{8\pi^3}\int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi} E^2[\Delta_I(\omega,\theta,\tau)]\mathrm{d}\omega\mathrm{d}\theta\mathrm{d}\tau.$$

where the sum is over all $(\delta_x, \delta_y, \delta_t) \in Q$. But the right-hand side of this identity is 0 by assumption. Thus, since each term in the left-hand sum is nonnegative, each must be 0 ∎

For current purposes, the importance of the Reichardt-detector characterization of the class of drift-balanced random stimuli (established in proposition 4) is that it provides

easy access to the principal results concerning the critical subclass of drift-balanced random stimuli that we call microbalanced. This is the focus of Section 8.

8. MICROBALANCED RANDOM STIMULI

Consider the following two-frame-block stimulus S: In frame block 0, a bright spot (call it spot 0) appears. In frame block 1, spot 0 disappears, and two new spots appear, one on each side of spot 0. On the one hand, it is clear (from proposition 4) that S is drift balanced. On the other hand, it is equally clear that a Fourier-based motion detector whose spatial reach encompassed the location of spot 0 and only one of the flashes in frame block 1 might be stimulated strongly in a fixed direction by S. Although S is drift balanced, some local Fourier motion detectors would be stimulated strongly and systematically by S. These detectors can be selected differentially by spatial windowing, and thereby a drift-balanced stimulus S can be converted into a non-drift-balanced stimulus.

In this section we introduce the class of microbalanced random stimuli, a subclass of drift-balanced random stimuli, any member I of which is guaranteed not to stimulate Fourier-power motion detectors in any systematic way, regardless of any space–time-separable window interposed between I and the detector. As we shall prove in proposition 8 below, I possesses this property if I satisfies the following definition.

Definition 6: Microbalanced Stimulus
Call any random stimulus I *microbalanced* iff, for any $(x, y, t), (x', y', t') \in Z^3$,

$$E[I[x, y, t]I[x', y', t']] = E[I[x, y, t']I[x', y', t]].$$

Obviously, for any random spatial function f and temporal random function g,

$$E[f[x, y]g[t]f[x', y']g[t']] = E[f[x, y]g[t']f[x', y']g[t]],$$

yielding the following proposition.

Proposition 5
Any space–time-separable random stimulus is microbalanced.

A related result is stated in the next proposition.

Proposition 6
Any invariant microbalanced stimulus I is space–time-separable.

Proof
If $I = 0$, there is nothing to prove (since, obviously, 0 is space–time separable). Otherwise we choose a point $(x', y', t') \in Z^3$, for which $I[x', y', t'] \neq 0$, and, for all $(x, y, t) \in Z^3$, we define

$$f(x, y) = I[x, y, t']$$

and

$$g(t) = \frac{I[x', y', t]}{I[x', y', t']}.$$

If either $(x, y) = (x', y')$ or $t = t'$, then immediately we obtain

$$I[x, y, t] = f(x, y)g(t).$$

On the other hand, if $(x, y) \neq (x', y')$ and $t \neq t'$, I's invariance and microbalancedness together imply that

$$I[x, y, t] = \frac{I[x, y, t']I[x', y', t]}{I[x', y', t']} = f(x, y)g(t). \quad \blacksquare$$

An important property of microbalanced random stimuli that sets them apart from the more general class of drift-balanced random stimuli is explained in proposition 7.

Proposition 7
The product of independent microbalanced random stimuli I and J is microbalanced.

Proof
For any $(x, y, t), (x', y', t') \in Z^3$,

$$\begin{aligned} E[IJ[x, y, t]IJ[x', y', t']] &= E[I[x, y, t]I[x', y', t']]E[J[x, y, t]J[x', y', t']] \\ &= E[I[x, y, t']I[x', y', t]]E[J[x, y, t']J[x', y', t]] \\ &= E[IJ[x, y, t']IJ[x', y', t]]. \quad \blacksquare \end{aligned}$$

Earlier in this section we showed, by using the example of a single spot splitting into two adjacent spots, that a drift-balanced random stimulus (S) can systematically stimulate motion detectors that operate on restricted regions of S. With proposition 8 we shall establish that all and only those random stimuli that are microbalanced avoid the systematic stimulation of all local (and global) Fourier-power detectors. The following lemma eases the proof of this important fact.

Lemma 3
Any microbalanced random stimulus is drift balanced.

Proof
Let I be microbalanced. From proposition 4, I is drift balanced iff $E[H_I[\delta_x, \delta_y, \delta_t]] = E[H_I[\delta_x, \delta_y, -\delta_t]]$ for any offset $(\delta_x, \delta_y, \delta_t) \in Z^3$, such that $(\delta_x, \delta_y) \neq (0, 0)$ and $\delta_t \neq 0$. However, since I is microbalanced, we note that for any such $(\delta_x, \delta_y, \delta_t)$,

$$\begin{aligned} E[H_I[\delta_x, \delta_y, \delta_t]] &= E[\sum I[x, y, t]I[x - \delta_x, y - \delta_y, t - \delta_t]] \\ &= \sum E[I[x, y, t]I[x - \delta_x, y - \delta_y, t - \delta_t]] \\ &= \sum E[I[x, y, t - \delta_t]I[x - \delta_x, y - \delta_y, t]] \\ &= E[\sum I[x, y, t - \delta_t]I[x - \delta_x, y - \delta_y, t]] \\ &= E[\sum I[x, y, t]I[x - \delta_x, y - \delta_y, t + \delta_t]] \\ &= E[H_I[\delta_x, \delta_y, -\delta_t]], \end{aligned}$$

where each of the sums is over all $(x, y, t) \in Z^3$. $\blacksquare$

We can now state the main result of this section.

Proposition 8
For any random stimulus I, the following conditions are equivalent:

I. I is microbalanced.

II. For any space–time-separable function W, WI is drift balanced.

Proof

First we prove that condition I implies condition II. Assume that I is microbalanced. By proposition 5, W is also microbalanced; it thus follows proposition 7 that WI is microbalanced and hence drift balanced (from lemma 3).

Next we prove that not condition I implies not condition II. Suppose that I is not microbalanced; then, for some (x, y, t), $(x', y', t') \in Z^3$,

$$E[I[x, y, t]I[x', y', t']] \neq E[I[x, y, t']I[x', y', t]].$$

[Note that this inequality implies that $(x, y) \neq (x', y')$ and $t \neq t'$.] Let f assign 1 to (x, y) and (x', y'), and let it assign 0 to all other points of Z^2; and let g assign 1 to t and t' and 0 to all other points of Z. Then the function fgI is zero everywhere except at the points (x, y, t), (x, y, t'), (x', y', t'), and (x', y', t'). It is obvious, from proposition 4, that fgI is not drift balanced. In particular,

$$\begin{aligned} E[H_{fgI}[x - x', y - y', t - t']] &= E[I[x, y, t]I[x', y', t']] \\ &\neq E[I[x, y, t']I[x', y', t]] \\ &= E[H_{fgI}[x - x', y - y', -(t - t')]]. \end{aligned}$$

∎

The results stated thus far in this section would not be interesting if there were no microbalanced random stimuli that displayed consistent apparent motion. The following result makes it clear that, in fact, all the examples of drift-balanced random stimuli that we considered previously are microbalanced.

Proposition 9

Let Γ be a family of pairwise independent, microbalanced random stimuli, all but at most one of which have an expectation of 0, then any linear combination of Γ is microbalanced.

Proof

Since a microbalanced random stimulus multiplied by a constant remains microbalanced, we assume that the linear combination is a sum; then, for any (x, y, t), $(x', y', t') \in Z^3$,

$$E\left[\sum_{I\in\Gamma} I[x, y, t] \sum_{J\in\Gamma} J[x', y', t']\right] = \sum_{I\in\Gamma}\sum_{J\in\Gamma} E[I[x, y, t]J[x', y', t']]. \quad (9)$$

However, whenever $I \neq J$,

$$E[I[x, y, t]J[x', y', t']] = E[I[x, y, t]]E[J[x', y', t']] = 0.$$

Thus Eq. (9) becomes

$$\sum_{I\in\Gamma} E[I[x, y, t]I[x', y', t']] = \sum_{I\in\Gamma} E[I[x, y, t']I[x', y', t]] = E\left[\sum_{I\in\Gamma} I[x, y, t'] \sum_{J\in\Gamma} J[x', y', t]\right].$$

∎

Next we secure the analog of proposition 2.

Proposition 10

The (spatiotemporal) convolution of two independent microbalanced random stimuli is microbalanced.

Proof

It is convenient to write

$$\sum_{\alpha,\beta,\ldots,\rho,\ldots}$$

for a sum in which each of the variables α, ranges over all integers. For any independent random stimuli I and J and any (x, y, t), $(x', y', t') \in Z^3$,

$$\begin{aligned} &E[I * J[x, y, t]I * J[x', y', t']] \\ &= E\left[\sum_{p,q,r} I[x - p, y - q, t - r]J[p, q, r] \times \sum_{p',q',r'} I[x' - p', y' - q', t' - r']J[p', q', r']\right] \\ &= \sum_{p,q,r,p',q',r'} E[I[x - p, y - q, t - r]I[x' - p', y' - q', t' - r']] \times E[J[p, q, r]J[p', q', r']]. \end{aligned}$$

But if, in addition, I and J are microbalanced, then this last sum is equal to

$$\begin{aligned} &\sum_{p,q,r,p',q',r'} E[I[x - p, y - q, t' - r']I[x' - p', y' - q', t - r]] \times E[J[p, q, r']J[p', q', r]] \\ &= E\left[\sum_{p,q,r'} I[x - p, y - q, t' - r']J[p, q, r'] \times \sum_{p',q',r} I[x' - p', y' - q', t - r]J[p', q', r]\right] \\ &= E[I * J[x, y, t']I * J[x', y', t]]. \end{aligned}$$

∎

Response of Reichardt Detectors to Microbalanced Random Stimuli

Two Fourier-analytic motion detectors proposed for psychophysical data[3,4] can be recast as variants of an ERD.[2,3] The ERD has many useful properties as a motion detector without regard to its specific instantiation.[1,2,21]

Figure 6 shows a diagram of the ERD. It consists of spatial receptors characterized by spatial functions f_1 and f_2, temporal filters g_1* and g_2*, multipliers, an adder, and another temporal filter $h*$. The spatial receptors f_i $(i = 1, 2)$ act on the input stimulus I to produce intermediate outputs,

$$y_i[t] = \sum_{(x,y)\in Z^2} f_i[x, y]I[x, y, t].$$

At the next stage, each temporal filter g_j* transforms its input y_i $(i, j = 1, 2)$, yielding four temporal output functions $g_j * y_i$. The left and right multipliers then compute

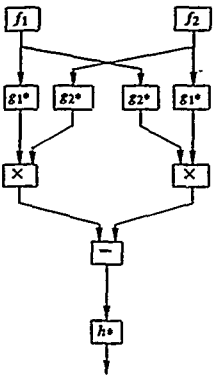


Fig. 6. Diagram of the ERD. Let I be a random stimulus, then, in response to I, for $i = 1, 2$, the box containing the spatial function $f_i: Z^2 \to \mathbb{R}$ outputs the temporal function $\sum_{(x,y)\in Z^2} f_i[x, y]I[x, y, t]$, each of the boxes marked g_i* outputs the convolution of its input with the temporal function $g_i: Z \to \mathbb{R}$; each of the boxes marked with a × outputs the product of its inputs, the box marked with a − outputs its left input minus its right, and the box containing h* outputs the convolution of its input with the temporal function $h: Z \to \mathbb{R}$.

$$[g_1 * y_1[t]][g_2 * y_2[t]], \qquad [g_1 * y_2[t]][g_2 * y_1[t]],$$

respectively, and the differencer subtracts the output of the right multiplier from that of the left multiplier:

$$D[t] = [g_1 * y_1[t]][g_2 * y_2[t]] - [g_1 * y_2[t]][g_2 * y_1[t]].$$

The final output is produced by applying the filter h*, whose purpose is to appropriately smooth the time-varying differencer output D.

In the following discussion, we write

$$\sum_{a_1, a_2, \ldots, a_n}$$

for a sum in which each of the variables a_i ranges over all integers. Given a random stimulus I as the input to the ERD, the output of the differencing component at time B is

$$D[B] = \left[\sum_u g_1[u] \sum_{x,y} f_1[x, y]I[x, y, B-u]\right] \times \left[\sum_t g_2[t] \sum_{p,q} f_2[p, q]I[p, q, B-t]\right] - \left[\sum_u g_1[u] \sum_{p,q} f_2[p, q]I[p, q, B-u]\right] \times \left[\sum_t g_2[t] \sum_{x,y} f_1[x, y]I[x, y, B-t]\right],$$

which can be rewritten as

$$D[B] = \sum_{t,u,p,q,x,y} g_1[u]g_2[t]f_1[x, y]f_2[p, q] \times [I[x, y, B-u]I[p, q, B-t] - I[x, y, B-t]I[p, q, B-u]].$$

However, if I is microbalanced, then (by definition 6) the expectation of the square-bracketed difference is 0, and hence $E[D[B]] = 0$ for any $B \in Z$, implying the following proposition.

Proposition 11
The expected response of any elaborated Reichardt detector to any microbalanced random stimulus is 0 at every instant in time.

Microbalanced random stimuli, then, compose a subclass of drift-balanced random stimuli with special importance for the investigation of non-Fourier motion perception. In general, the fact that a random stimulus I is drift balanced does not entail that all local areas of I be drift balanced; that is, the window over which the Fourier power analysis of I is carried out is critical to the drift-balancedness of I. This constraint is escaped by microbalanced random stimuli (as a consequence of proposition 8). a random stimulus I is microbalanced iff, for any space-time-separable function W, the random stimulus WI (the result of windowing I by W) is drift balanced.

9. RECOVERY OF MOTION FROM MICROBALANCED RANDOM STIMULI

Nonlinear Transformations Hypothesis

The most plausible explanation for the recovery of motion from drift-balanced random stimuli posits one or more nonlinear transformations that are routinely applied to the visual input signal to generate a new signal, which is then subjected to ordinary frequency-domain power analysis.

Consider, for instance, random stimuli such as those described in demonstrations 1, 2, and 5 (Figs. 4a, 4b, and 4e), whose motion depends on spatiotemporal modulation of noise contrast. For concreteness, we focus on I, the contrast-reversing bar of demonstration 1 (Fig. 4a). The apparent motion exhibited by I might result from a power analysis in the frequency domain of a rectified version of the original signal for example, a transformation of the signal I such as R_I, S_I, T_I^+, or T_I^-, where

(i) $R_I[x, y, t] = |I[x, y, t]|$ (full-wave rectification),
(ii) $S_I[x, y, t] = I[x, y, t]^2$ (full-wave power rectification),
(iii) $T_I^+[x, y, t] = \max\{I[x, y, t], 0\}$ (positive half-wave rectification),
(iv) $T_I^-[x, y, t] = \min\{I[x, y, t], 0\}$ (negative half-wave rectification).

R_I and S_I both transform I into a rectangle moving in a series of brief steps from left to right, while T_I^+ and T_I^- map I into a similar such moving rectangle, which randomly disappears and reappears in the course of its left-to-right traversal. The MFFC principle applied to any of these transformations of I would indicate motion to the right (see Fig 7a). In the

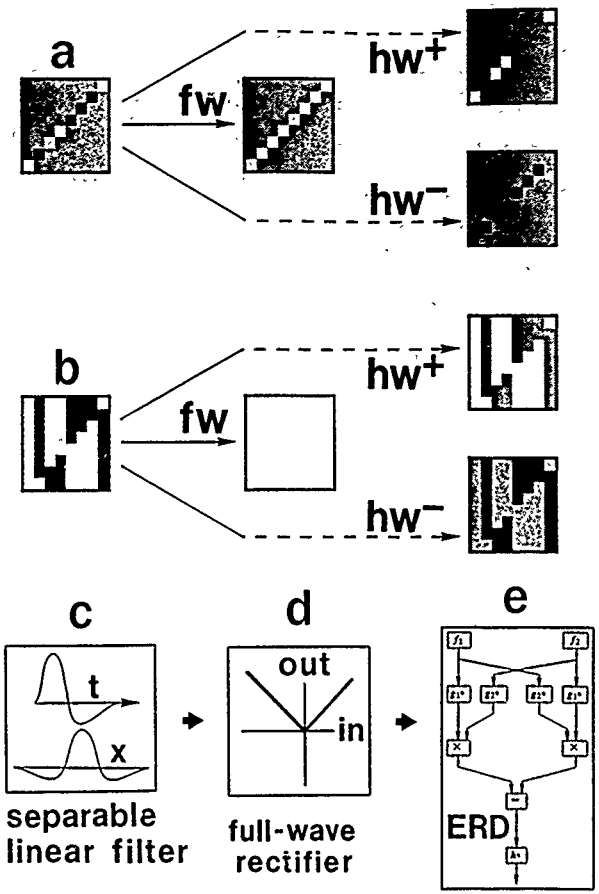


Fig 7 Consequences of full-wave and half-wave rectification. a, Space–time representation of a traveling, contrast-reversing bar; full wave (fw) rectified representation, and positive (hw^+) and negative (hw^-) half wave rectified representations, showing that either of these rectifications suffices to expose the motion to Fourier motion energy analysis. b, Space–time representation of a traveling contrast reversal of a random bar pattern, full wave (fw) rectified representation, positive (hw^+) and negative (hw^-) half-wave rectified representations, showing that none of these rectifications exposes motion. The analysis system for second order motion stimuli is shown in the bottom row. c, the signal is linearly filtered (the impulse response of an appropriate space–time-separable linear filter is shown), d, the filtered signal is full wave rectified, and e, it is subjected to motion energy analysis (e g , by an ERD). This is a sufficient sequence of operations to expose the directional motion in all the demonstrations of this paper.

realm of spatial visual perception, rectification transformations were proposed by various authors to mediate boundary formation and texture segregation.[24-28] Logarithmic intensity compression was also proposed,[29-32] because of its physiological plausibility, although it is less effective than rectification.

Although any one of the rectification transformers would expose the motion information buried in I to frequency-domain power analysis, the same is not true of the traveling contrast-reversal J defined in demonstration 3 (Fig 4c). Full-wave rectification of J yields a constant output. Half-wave rectification merely yields another drift-balanced random stimulus $T_J^+ = (J + 1)/2$ and $T_J^- = (1 - J)/2$. These relations are illustrated in Fig 7b. The motion of J does not emerge directly from any of these forms of rectification.

For the traveling, random contrast-reversal J (demonstration 3, Fig. 4c), a time-dependent linear operator such as temporal differentiation is required to transform it into a

signal from which motion information can be extracted after rectification. (Indeed, the partial derivative of J with respect to time is I.)

Consider the space-time-separable bandpass filtering that is usually assumed to occur in low-level visual processing. If such linear filtering were applied to any of the demonstrations considered in this paper, and if it were followed by any of the rectification operations considered above, it would suffice to expose the motion of any of these demonstrations to Fourier power analysis. Figure 7 illustrates the sequence of filtering, rectifying, and motion-power analysis. A central issue concerning drift-balanced random stimuli thus emerges: given the (largely unexplored) range of drift-balanced random stimuli that elicit apparent motion, what is the simplest array of transformations of the input signal that suffices to expose (to frequency-domain power analysis) the motion information carried by all the various types of drift-balanced random stimuli?

What Is the Purpose of Having Detectors for Drift-Balanced Motion?

From a systems point of view, there is a problem in linearly combining the information from many linear sensors (for example, motion-sensitive sensors) because there is nothing gained by the combination that could not have been accomplished by a single, large sensor. For an advantage to be gained from the combination, this information must be nonlinearly related to the input. Nonlinearly computed quantities such as power and information are combined most usefully. In many classical detection problems the ideal detector is a power detector; that is, the power of the component elements is summed to form the decision variable.[33,34] When it comes to detecting motion, it would be surprising if generally similar considerations did not apply in combining information from various locations of the visual field and from detectors of various sizes. Indeed, the MFFC theories normally use motion detectors that compute Fourier power.[1-6]

Assuming that evolution chooses detection modes because of their advantages, what is surprising about the detection of drift-balanced motion is that the advantages of nonlinear combination are already available at the earliest stages of sensory analysis. Ultimately, to appreciate why this is so requires ecological analysis of the visual world. Obviously, the ecological problem cannot be resolved by armchair speculation. On the other hand, given that combination mechanisms operate with rectified inputs, it is not surprising that the mechanisms that detect drift-balanced motion seem to be of a much larger scale than the Fourier mechanisms.[35] A possibly related observation is that the apparent motion in various drift-balanced random stimuli that we have considered here tends to diminish with the retinal eccentricity of the presentation.[11] However, it remains to be determined how much of this drop-off of apparent motion should be attributed to the effective decrease in visual spatial sampling rate with retinal eccentricity.

10. UTILITY OF RANDOM STIMULI AS A RESEARCH TOOL

A general advantage of random stimuli compared with repeated stimuli is that the responses to a repeated stimulus might be mediated by any of its features, including artifactual stimulus features that are not anticipated by the experimenter. Responses to random stimuli represent the responses to the properties that distinguish a class of stimuli, and these tend to be more general and more readily specifiable than the properties of a single stimulus. Thus, by generalizing the notion of a stimulus to that of a random stimulus, we obtain a much more extensive and adaptable set of tools for studying perception.

In the study of motion perception, microbalanced random stimuli play a crucial role: they avoid the complications introduced by the spatial windowing that is unavoidably performed by motion-perception units. Avoiding the possible artifacts of windowing is particularly important in interpreting the responses of single visual neurons. Only a microbalanced random stimulus is guaranteed to contain no consistent Fourier components, regardless of how that stimulus may be centered or fail to be centered in a given neuron's receptive field or in the observer's field of view. It is possible for drift-balanced (but not microbalanced) random stimuli to produce systematic Fourier motion components in receptive fields of particular neurons that happen to be placed advantageously with respect to those stimuli. Only microbalanced random stimuli necessarily require non-Fourier operations in order to yield motion perception.

An invariant stimulus is microbalanced (thereby avoiding the windowing problem) only if it is space-time separable (proposition 6). Unfortunately, there are no examples of space-time-separable stimuli that yield a strong, consistent perception of motion. Thus random microbalanced stimuli that yield strong perceived motion offer a unique tool for the investigation of non-Fourier motion perception.

11. NON-FOURIER STIMULUS ANALYSIS IN OTHER SENSORY DOMAINS

Spatial Vision

One-dimensional motion stimuli in (x, t) can be represented as two-dimensional stimuli in (x, y). From the point of view of systems analysis, the (x, t) and (x, y) representations are equivalent: motion in (x, t) is equivalent to orientation in (x, y). There are inevitably some physical restrictions that apply in the time domain,[2] so that x and t cannot be so symmetrical with respect to each other as x and y. For example, in human motion detectors, summation over time (of comparator output) occurs within a single detector, summation over space occurs between detectors.

The space–time asymmetry in motion can be made obvious by adding two gratings. Thus, when a drifting sine-wave grating of frequency (ω_x, ω_t) is added to a stationary sine pattern of frequency $(\omega_x, 0)$ (a standing grating), the apparent motion is normally visible, when it is added to $(0, \omega_t)$ (a uniform, flickering field), the apparent motion may either be normal or be reversed, depending on the phase relations.[2] In the space domain, both combinations are equivalent.

The fact that all the (x, y) spatial illustrations in the figures of (x, t) motions were visible as oriented textures demonstrates that the same or similar nonlinear dynamics are involved in the extraction of orientation as are involved in the extraction of direction of motion. Indeed, we have yet to discover an (x, t) stimulus that is perceived as moving and that is not perceived as oriented texture in an (x, y)

representation. This suggests that the human array of pattern-analytic detectors is at least as rich as the motion-analytic array.

Audition

Obviously, a one-dimensional signal, such as an auditory signal (which depends only on time), cannot be drift balanced. Nonetheless, certain auditory phenomena bear a resemblance to some of the visual effects that we have been considering.

It has long been recognized that the auditory system analyzes sound-pressure waveforms into their component sinusoidal frequencies and that these frequency components correspond, at least to a first approximation, to the sensation of pitch. Indeed, the cochlea functions largely as a mechanical frequency analyzer. In addition to pure frequency analysis, especially at periodicities below 300 Hz, another mechanism, periodicity analysis, also comes into play. One of the best demonstrations is an experiment by Miller and Taylor.[36]

Some background facts about this experiment are useful here. A broad-spectrum noise N is a random function of time such that the expected power of all Fourier components in N is equal. It is easy to show that any random function N that assigns pairwise independent random variables, all with mean 0, to distinct points in time is a broad-spectrum noise. Obviously, multiplying any such random function N by an arbitrary nonrandom function f yields yet another broad-spectrum noise, since the values assigned by fN remain pairwise independent, each with mean 0.

In the experiment by Miller and Taylor, listeners heard a broad-spectrum noise that was modulated on and off (multiplied) by a square wave of frequency f. Thus the stimulus generated by Miller and Taylor had a uniform expected power over all temporal frequencies. When f was less than ~10 Hz, the perception corresponded to the physical reality of interrupted noise. At frequencies between 40 and 200 Hz, the interrupted noise was perceived to have a pitch that corresponded to the interruption frequency. That observers perceive a pitch implicates some mechanism other than frequency analysis. Whereas a rectifying nonlinearity was not proposed explicitly by Miller and Taylor, it is the obvious intermediate step in periodicity pitch perception.

12. FINAL REMARKS

We have given precise definition to the notion of a random stimulus and focused our attention on the subclasses of drift-balanced and microbalanced random stimuli as being especially interesting for the study of visual perception. We first showed that the (spatiotemporal) convolution of independent drift-balanced random stimuli is drift balanced.

Proposition 3 (which states that the sum of drift-balanced random stimuli is drift balanced when the elements are pairwise independent and all but at most one have expectation 0, the non-0 element being invariant) and proposition 9 (which states a similar result for microbalanced random stimuli) provide access to a large family of empirically useful drift-balanced random stimuli. Instances that display striking apparent motion may be constructed readily.

In Section 8 we introduced microbalanced random stimuli, a distinguished subclass of drift-balanced random stimuli defined by the following property. A random stimulus I is microbalanced iff, for any space-time-separable function W, the product WI is drift balanced. Thus I is guaranteed to avoid systematically stimulating any Fourier power motion mechanisms encountering I through any space–time-separable window. It was proved that (proposition 5) any space–time-separable random stimulus is microbalanced; that (proposition 6) any invariant microbalanced stimulus is space–time separable; that (proposition 7) the product of two independent microbalanced random stimuli is microbalanced; that (proposition 9) any linear combination of pairwise independent microbalanced random stimuli, all but at most one of which has expectation 0, is microbalanced; and that (proposition 10) the spatiotemporal convolution of two independent microbalanced random stimuli is microbalanced. An implication of proposition 9 is that all the demonstration stimuli presented in this paper are not only drift balanced but also microbalanced. Finally (in proposition 11), we showed that the expected response of any elaborated Reichardt detector to any microbalanced random stimulus is 0 at any instant in time.

In light of earlier observations,[7-14] the existence of non-Fourier mechanisms is hardly surprising. Such mechanisms have, however, received no thorough investigation. The range of types of such mechanisms has not yet been elaborated, and their psychophysical properties remain largely unstudied. The importance of proposition 3 and the results of Section 8 lies in their utility for constructing stimuli for probing both the nature of non-Fourier motion-detection mechanisms as well as the interaction between such mechanisms and the band-tuned motion detectors that were the focus of most previous research.

APPENDIX A

In this appendix we verify that $E[|\tilde{I}(\omega, \theta, \tau)|^2]$ exists for any random stimulus I and any $\omega, \theta, \tau \in \mathbb{R}$ (which was presumed in definition 2). Let $D = \{(x, y, t) \in \mathbb{Z}^3 | I[x, y, t] \neq 0\}$, then

$$E[|\tilde{I}(\omega, \theta, \tau)|^2] = \int_{\mathbb{R}^{\#D}} \sum i[x, y, t] i[p, q, r] \times \exp\{j(\omega(x - p) + \theta(y - q) + \tau(t - r))\} f(i) \mathrm{d}i$$

$$= \sum \int_{\mathbb{R}^{\#D}} i[x, y, t] i[p, q, r] f(i) \mathrm{d}i \times \exp\{j(\omega(x - p) + \theta(y - q) + \tau(t - r))\},$$

where each sum ranges over all pairs of points, (x, y, t), $(p, q, r) \in \mathbb{Z}^3$. Note now that

$$\int_{\mathbb{R}^{\#D}} i[x, y, t] i[p, q, r] f(i) \mathrm{d}i = E[I[x, y, t] I[p, q, r]].$$

However, as a consequence of the (probabilistic version of the) Schwartz inequality,[37] we note that

$$E[I[x, y, t] I[p, q, r]] \leq (E[I[x, y, t]^2] E[I[p, q, r]^2])^{1/2}.$$

However, by the definition of a random stimulus, the two expectations on the right-hand side of the inequality exist. Hence $E[|\tilde{I}(\omega, \theta, \tau)|^2]$ exists for all $\omega, \theta, \tau \in \mathbb{R}$.

APPENDIX B

In this appendix we prove lemma 1, which is as follows:

Let S be a random stimulus equal to the sum of a set Ω of pairwise independent random stimuli; then

$$E[|S|^2] = |E_S|^2 + \sum_{I \in \Omega} E[|\hat{N}_I|^2],$$

where $\hat{N}_I = I - E_I$ for each $I \in \Omega$.

First we write

$$S = \sum_{I \in \Omega} (E_I + N_I).$$

The linearity of Fourier transformation then yields

$$\hat{S} = \sum_{I \in \Omega} (\hat{E}_I + \hat{N}_I).$$

Thus

$$|\hat{S}|^2 = \Sigma[\hat{E}_I(\hat{E}_J)^* + \hat{N}_I(\hat{N}_J)^* + \hat{E}_I(\hat{N}_J)^* + \hat{N}_I(\hat{E}_J)^*],$$

where the sum is over all $I, J \in \Omega$.

Note first, however, that, whenver $I \neq J$,

$$E[\hat{E}_I(\hat{E}_J)^* + \hat{N}_I(\hat{N}_J)^* + \hat{E}_I(\hat{N}_J)^* + \hat{N}_I(\hat{E}_J)^*] = \hat{E}_I(\hat{E}_J)^*,$$

since I and J are independent and

$$E[\hat{N}_I] = E[(\hat{N}_J)^*] = \hat{0}.$$

Moreover, whenever $I = J$,

$$E[\hat{E}_I(\hat{E}_I)^* + \hat{N}_I(\hat{N}_I)^* + \hat{E}_I(\hat{N}_I)^* + \hat{N}_I(\hat{E}_I)^*] = \hat{E}_I(\hat{E}_I)^* + E[\hat{N}_I(\hat{N}_I)^*].$$

Thus

$$\begin{aligned} E[|\hat{S}|^2] &= \sum_{J \in \Omega} \sum_{K \in \Omega} \hat{E}_J(\hat{E}_K)^* + \sum_{I \in \Omega} E[|\hat{N}_I|^2] \\ &= \left| \sum_{J \in \Omega} \hat{E}_J \right|^2 + \sum_{I \in \Omega} E[|\hat{N}_I|^2] \\ &= |E_S|^2 + \sum_{I \in \Omega} E[|\hat{N}_I|^2]. \end{aligned}$$

∎

APPENDIX C

In this appendix we prove that the random stimuli G and H of demonstrations 5 and 4 are drift balanced These random stimuli stem from proposition 3. To make the bridge explicit, we shall need to derive a corollary (C1) that depends on the following lemma.

Lemma C1

For $M \in \mathbf{Z}^+$, let the random variables $\rho_0, \rho_1, \ldots, \rho_{M-1}$ be pairwise independent, each uniformly distributed on $[-\pi, \pi)$, then, for any $x, y, t \in \mathbf{Z}$, define the random stimulus I by setting

$$I[x, y, t] = \sum_{m=0}^{M-1} d_m[x, y] (\cos(\rho_m) h_m[t] - \sin(\rho_m) k_m[t]),$$

where, in each case, d_m, h_m, and k_m are all real-valued functions that equal zero at all but a finite number of points of their respective domains. I is then drift balanced.

Proof

For $m = 0, 1, \ldots, M-1$, term m of I is space–time separable and hence drift balanced. Moreover, for each m, the expectations of $\sin(\rho_m)$ and $\cos(\rho_m)$ are both 0. Thus the expectation of each term of the sum yielding I is 0; the result follows from proposition 3. ∎

We apply lemma C1 to prove the following corollary used in constructing stimuli for demonstrations 4 and 5.

Corollary C1

For $M, N \in \mathbf{Z}^+$, let $\rho_0, \rho_1, \ldots, \rho_{M-1}$ be pairwise independent random variables, each uniformly distributed on $[-\pi, \pi)$; then, for any $x, y, t \in \mathbf{Z}$, define the random stimulus I by setting

$$I[x, y, t] = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} d_m[x, y] p_{m,n}[t] \cos(q_{m,n}[t] + \rho_m),$$

where, for $m = 0, 1, \ldots, M-1$, and $n = 0, 1, \ldots, N-1$, the functions d_m, $p_{m,n}$, and $q_{m,n}$ are real valued and zero at all but a finite number of points of their corresponding domains. I is then drift balanced.

Proof

We recast I so as to apply lemma C1:

$$\begin{aligned} I[x, y, t] &= \sum_{m=0}^{M-1} d_m[x, y] \sum_{n=0}^{N-1} p_{m,n}[t] \\ &\quad \times (\cos(q_{m,n}[t]) \cos(\rho_m) - \sin(q_{m,n}[t]) \sin(\rho_m)) \\ &= \sum_{m=0}^{M-1} d_m[x, y] (h_m[t] \cos(\rho_m) - k_m[t] \sin(\rho_m)) \end{aligned}$$

for

$$h_m[t] = \sum_{n=0}^{N-1} p_{m,n}[t] \cos(q_{m,n}[t]),$$

$$k_m[t] = \sum_{n=0}^{N-1} p_{m,n}[t] \sin(q_{m,n}[t]).$$

∎

Proof That H (Demonstration 4) Is Drift Balanced

H contains N frame blocks indexed $0, 1, \ldots, N-1$, each composed of M rectangles indexed $0, 1, \ldots, M-1$ from left to right. Let $\rho_0, \rho_1, \ldots, \rho_{M-1}$ be pairwise independent random variables, each uniformly distributed on $[-\pi, \pi)$. Let C be a contrast value. We can express H as follows. For $m = 0, 1, \ldots, M-1$, let $d_m[x, y] = 1$ for (x, y) in the mth rectangle and 0 elsewhere, and for $n = 0, 1, \ldots, N-1$, let $g_n[t] = 1$ in the nth frame block and 0 elsewhere; then

$$H[x,y,t] = C\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} d_m[x,y]g_n[t]$$

$$\times\cos\left(4\pi\left(1+\cos\left(2\pi\left(\frac{m}{M}-\frac{n}{N}\right)\right)\right)+\rho_m\right).$$

To check that H is drift balanced, make the following identifications, and then apply corollary C1:

$$p_{m,n}[t] = Cg_n[t]$$

and

$$q_{m,n}[t] = 4\pi\left(1+\cos\left(2\pi\left(\frac{m}{M}-\frac{n}{N}\right)\right)\right).$$

Thus corollary C1 applies, and we conclude that H is drift balanced. (Note that H does not exploit the full generality of corollary C1, since, for these identifications, $p_{m,n}[t]$ does not depend on m and $q_{m,n}[t]$ does not depend on t.)

Proof That G (Demonstration 5) Is Drift Balanced

The random stimulus G is made up of N frame blocks indexed $0, 1, \ldots, N-1$, each containing M rectangles indexed $0, 1, \ldots, M-1$ from left to right. Let $\rho_0, \rho_1, \ldots, \rho_{M-1}$ be pairwise independent random variables, each uniformly distributed on $[-\pi, \pi)$. Let C be some contrast value. We can then express G as follows. For $m = 0, 1, \ldots, M-1$, let $d_m[x, y] = 1$ for (x, y) in the mth rectangle and 0 elsewhere; for $n = 0, 1, \ldots, N-1$, let $g_n[t] = 1$ for t in the nth frame block and 0 elsewhere; then

$$G[x,y,t] = \frac{C}{2}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} d_m[x,y]g_n[t]$$

$$\times\left(\cos\left(2\pi\left(\alpha\frac{m}{M}-\beta\frac{n}{N}\right)\right)+1\right)\cos\left(2\pi\gamma\frac{n}{N}+\rho_m\right).$$

To see that G is drift balanced, set

$$p_{m,n}[t] = \frac{C}{2}g_n[t]\left(\cos\left(2\pi\left(\frac{\alpha m}{M}-\frac{\beta n}{N}\right)\right)+1\right)$$

and

$$q_{m,n}[t] = 2\pi\frac{\gamma n}{N},$$

and apply corollary C1. (Note that, as with H, G does not exploit the full generality of corollary C1, since $q_{m,n}[t]$ depends on neither m nor t.)

ACKNOWLEDGMENTS

The research reported here was supported by U.S. Air Force Life Science Directorate, Visual Information Processing Program, grant 85-0364. The authors thank Michael S. Landy for helpful comments on various drafts of the manuscript.

REFERENCES

1. J. P. H. van Santen and G. Sperling, "Temporal covariance model of motion perception," J Opt. Soc. Am A 1, 451–473 (1984)
2. J. P. H. van Santen and G. Sperling, "Elaborated Reichardt detectors," J. Opt. Soc. Am. A 2, 300–321 (1985).
3. E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," J. Opt. Soc. Am. A 2, 284–299 (1985).
4. A. B. Watson and A. J. Ahumada, "A look at motion in the frequency domain," NASA Tech. Memo. 84352 (National Aeronautics and Space Administration, Washington, D.C., 1983)
5. D. J. Fleet and A. D. Jepson, "On the hierarchical construction of orientation and velocity selective filters," Tech. Rep RBCV-TR-85-8 (Department of Computer Science, University of Toronto, Toronto, 1985).
6. D. J. Heeger, "Model for the extraction of image flow," J. Opt. Soc. Am. A 4, 1455–1471 (1987).
7. A. Pantle and L. Picciano, "A multistable movement display: evidence for two separate motion systems in human vision," Science 193, 500–502 (1976).
8. M. Green, "What determines correspondence strength in apparent motion," Vision Res. 26, 599–607, 1986.
9. A. M. Derrington and G. B. Henning, "Errors in direction-of-motion discrimination with complex stimuli," Vision Res. 27, 61–75 (1987).
10. A. M. Derrington and D. R. Badcock, "Separate detectors for simple and complex grating patterns?" Vision Res. 25, 1869–1878 (1985).
11. A. Pantle and K Turano, "Direct comparisons of apparent motions produced with luminance, contrast-modulated (CM), and texture gratings," Invest. Ophthalmol. Vis Sci. 27, 141 (1986).
12. K. Turano and A. Pantle, "On the mechanism that encodes the movement of contrast variations. I. velocity discrimination," submitted to Vision Res.
13. G. Sperling, "Movement perception in computer-driven visual displays," Behav. Res Methods Instrum 8, 144–151 (1976).
14. J. T. Petersik, K I. Hicks, and A. J. Pantle, "Apparent movement of successively generated subjective figures," Perception 7, 371–383 (1978).
15. C. Chubb and G. Sperling, "Drift-balanced random stimuli: a general basis for studying non-Fourier motion perception," Invest. Ophthalmol. Vis. Sci. 28, 233 (1987).
16. The main demonstrations and results described herein were first reported at the Symposium on Computational Models in Vision, Center for Visual Science, University of Rochester, June 20, 1986, and at the meeting of the Association for Research in Vision and Opthalmology, Sarasota, Florida, May 7, 1987.
17. A. B. Watson and A. J. Ahumada, Jr., "Model of human visual-motion sensing," J. Opt. Soc. Am. A 2, 322–342 (1985).
18. J. Victor, "Nonlinear processes in spatial vision: analysis with stochastic visual textures," Invest. Ophthalmol Vis Sci. 29, 118 (1988).
19. C. Chubb and G. Sperling, "Texture quilts. basic tools for studying motion from texture," Publ. 88-1 of *Mathematical Studies in Perception and Cognition* (Department of Psychology, New York University, New York, 1988).
20. W. Reichardt, "Autokorrelationsauswertung als Funktionsprinzip des Zentralnervensystems," Z. Naturforschung Teil B 12, 447–457 (1957).
21. J P. H van Santen and G Sperling, "Applications of a Reichardt-type model of two-frame motion," Invest. Ophthalmol. Vis. Sci. 25, 14 (1984).
22. A. B Watson, A J. Ahumada, and J. E Farrell, "The window of visibility: a psychophysical theory of fidelity in time sampled motion displays," NASA Tech Paper 2211 (National Aeronautics and Space Administration, Washington, D C., 1983)
23. A B Watson, A. J. Ahumada, Jr , and J E. Farrell, "Window of visibility: a psychophysical theory of fidelity in time-sampled visual motion displays," J. Opt Soc Am A 3, 300–307 (1986).
24. Grossberg S and E Mingolla, "Neural dynamics of form perception boundary completion, illusory figures and neon color spreading," Psychol. Rev. 92, 173–211 (1985)
25. S. Grossberg and E. Mingolla, "Neural dynamics of form perception textures, boundaries, and emergent segmentations," Percept Psychophys 38, 141–171 (1985).
26. R J Watt and M J. Morgan, "The recognition and representation of edge blur evidence of spatial primitives in human vision," Vision Res. 23, 1465–1477 (1983).

Spatial-frequency bands in complex visual stimuli: American Sign Language

Thomas R. Riedl* and George Sperling

Human Information Processing Laboratory, Department of Psychology, New York University, New York, New York 10003

Received April 27, 1987; accepted December 19, 1987

Dynamic images of individual signs of American Sign Language (ASL) with a resolution of 96 × 64 pixels were bandpass filtered in adjacent frequency bands. Intelligibility was determined by testing deaf subjects fluent in ASL. The following results were obtained. (1) By iteratively varying the center frequencies and bandwidths of the spatial bandpass filters, it was possible to divide the original signal into four different component bands of high intelligibility. (2) The measured temporal-frequency spectrum was approximately the same in all bands. (3) The masking of signals in band i by noise in band j was found to be inversely proportional to $\log |f_{signal}/f_{band}|$. At constant performance, the ratio of root-mean-square signal amplitude to noise amplitude, s/n, was the same for bands 2, 3, and 4 and higher for band 1. (4) When weak signals i and j were added linearly, there was a slight intelligibility advantage for signals in the same band ($i = j$) compared with signals in adjacent bands and for signals in adjacent bands compared with signals in distant bands.

INTRODUCTION

Much has been learned about how the spatial-frequency components of simple visual stimuli, in combination, contribute to visual responses. Most of what we know is concerned with simple stimuli near their threshold.[1] For example, there is ample evidence that multiple channels (mechanisms) are involved in the detection of simple visual stimuli—different channels at different retinal spatial frequencies.[2] It is believed that, at threshold, these channels sum their information probabilistically. Whether a channel that subserves one spatial frequency inhibits channels that subserve other frequencies is unclear; different results are reported for different procedures.[1]

Much of the visual research is concerned with spatial frequencies as they are produced at the retina. The discriminability of stimuli that are well above threshold, and explicitly limited by external noise, is independent of viewing distance (retinal angle) over a wide range.[3,4] Noisy signals are discriminated equally at vastly different retinal frequencies, and their perceptual properties are best characterized by cycles per object rather than cycles per degree of visual angle.

In a visual communication channel for complex, dynamic visual stimuli, such as American Sign Language (ASL), the limitations are related to stimulus noise and to stimulus subsampling rather than to low contrast, that is, the intelligibility of these ASL stimuli is limited by external distortions, modeled as noise, rather than by internal noise. Such limitations will probably be characterized by object spatial frequencies,[5] and almost none of the previous literature on spatial-frequency interactions in vision is directly applicable. Therefore, to design optimal communication channels for transmitting dynamic complex stimuli, there is no alternative to studying them directly.

From a practical point of view, visual communication channels would be immediately useful to the several hundred thousand hearing-impaired individuals who rely on ASL for communication.[6] More than two million Americans are unable to understand speech even with a hearing aid; many of these would benefit by having a visual communication channel to aid their utilization of residual hearing. The problem is that available, affordable channel capacity is limited, and compressing images to utilize this capacity efficiently requires a better understanding of how frequency components of complex images contribute to their intelligibility as well as better methods of image compression.[7-9]

This study is concerned with how the visual information in component spatial-frequency bands of a complex visual signal, ASL, combines to facilitate or to interfere with the intelligibility of ASL. Therefore first we attempt to establish four spatial-frequency bands having approximately equal intelligibility for ASL. Second, we measure the temporal characteristics of each of these bands. Third, we study how various intensities of noise in frequency band i interfere with signals in band j. Fourth, we determine how weak signals in band i combine with weak signals in band j to facilitate perception.

EXPERIMENT 1: BANDS OF EQUAL INTELLIGIBILITY

The purpose of experiment 1 is to derive a number of spatial-frequency filters to produce bandpass ASL stimuli from the original ASL stimuli. Each band should have approximately equal, and moderately high, intelligibility. Preliminary work suggested that four such bands would be possible for our stimuli.

Method

Original Stimuli

The stimuli consisted of isolated ASL signs displayed at 30 frames per second (fps) on a television raster monitor. Signs took 2–3 sec and consisted of 60–90 frames. A stan-

0740-3232/88/040606-11$02.00 Atch-34

Fig. 2. The ASL images filtered in bands 1–4. The leftmost image is the unfiltered original.

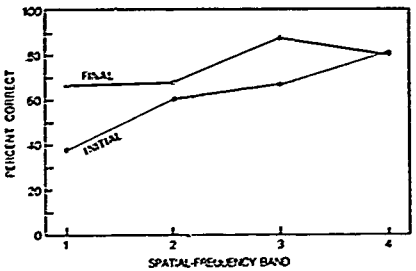


Fig. 3. Intelligibility (percentage of correct ASL sign identifications) as a function of the spatial-frequency band. Curve labeled INITIAL was obtained in experiment 1a with the filter set at top of Fig. 1, curve labeled FINAL was obtained in experiment 1b with filters at the bottom of Fig. 1 and with improved stimuli.

procedure by a proficient signer. The signs were run in blocks (by frequency band) so that the signer would be maximally prepared for the type of stimulus to be shown on a trial.

Results

The average percentages of correct responses in each band are shown in Table 1. As can be seen, performance improves with increasing frequency, from 38% in band 1 to 80% in band 4.

Experiment 1b: Filter Set 3

Procedure

Filter set 1 did not generate equally intelligible bands. Therefore the filters were changed according to an algorithm that estimated the contribution to intelligibility of every component frequency and attempted to distribute these contributions equally among the bands. In addition to intelligibility differences among bands in experiment 1a, we noted that there were some unfamiliar signs and that these may not have been distributed equally among groups. Therefore, for subsequent tests, 28 ambiguous signs were discarded. The remaining 72 signs were divided into four groups and were tested as before. Subsequently, the filters were again adjusted by an algorithm to increase the bandwidth of the bands with the worst performance and to diminish the bandwidth of the bands with the best performance. The final filters are shown in Fig. 1, and examples of the filtered stimuli are illustrated in Fig. 2.

To make the intelligibility test more accurate, data collected up to this point were used to rank the signs into three categories: easy, medium, and difficult. Signs in each category were distributed evenly into band conditions. Further, a balanced Latin square block design was used so that each sign was processed in each frequency band; i.e., four complete stimulus video tapes were prepared, each of which contained all the experimental ASL signs but distributed into different filter groups. Eight subjects were run, two subjects for each cell of the Latin square.

Results

Filter set 3 yielded four bands with intelligibilities that were more nearly equal than those of filter set 1, but intelligibility was still not completely uniform across bands. Intelligibility ranged from 66% in band 1 to 87% in band 3 (Fig. 3). Although the four bands of filter set 3 were not equally intelligible, they were sufficiently close to equal that we could move forward with the main experiments to investigate how signals in different bands interfere with and facilitate one another.

EXPERIMENT 2: THE TEMPORAL-FREQUENCY SPECTRUM

Here we address the question: What is the temporal power spectrum of the signal in each of the spatial bands derived in experiment 1? This question is of interest in its own right in terms of discovering the correlation of spatial and temporal frequencies in the environment and therefore in defining the optimal visual detectors for operating in this environment. More immediately, we will need the temporal data in experiment 3 to create dynamic visual noise that is matched to the spatially band-limited ASL signals in both spatial and temporal frequency.

To determine the signal power as a function of temporal frequency, eight representative ASL signs were selected. At the mean spatial frequency m_i of each spatial-filter i of experiment 1 (see Table 1, column 6), a small spatial-frequency range Δm_i [$\Delta m_i = \{(\omega_x,\omega_y)|m^2 - \epsilon \le \omega_x^2 + \omega_y^2 < m^2 + \epsilon\}$] was selected for analysis. This is the range of spatial frequencies that best characterizes its spatial-frequency band.

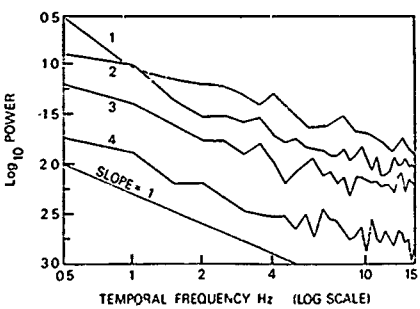


Fig. 4. The temporal power spectrum of ASL in spatial frequency bands 1–4. The abcissa represents the temporal frequency in hertz, the maximum frequency of 15 Hz is determined by the frame rate of 30 Hz. The ordinate represents the average power in an annular band of temporal frequencies extracted from a three-dimensional (x, y, t) Fourier analysis of eight representative ASL sign sequences. The line of slope −1 is drawn for reference.

The spatial range Δm_i is an annulus in ω_x–ω_y spatial-frequency space and a hollow cylinder in $(\omega_x, \omega_y, \omega_t)$ spatiotemporal-frequency space. For every small range of temporal frequencies Δf within Δm_i, the average power (over the eight signs) was computed at each spatiotemporal frequency (annular cross section of the cylinder). The whole computation was repeated for each of four spatial bands i. These data (temporal power versus temporal frequency, for each of the four spatial frequencies m_i) are displayed in Fig. 4.

Overall temporal power diminishes with increasing spatial frequency. Within each spatial-frequency band, temporal power falls off with an initial slope of approximately -1 on the graph of $\log_{10}$ (power) versus $\log_{10}$ (frequency), leveling off at high temporal frequencies. The approximate parallelism of the temporal-frequency power curves (for different spatial frequencies) suggests that the temporal-frequency composition of our ASL stimuli is independent of their spatial composition.

EXPERIMENT 3: CROSS-BAND MASKING BY NOISE

Typically, cross-band masking has been studied with simple static signals[1,2,12,18] rather than with realistic dynamic stimuli. The purpose of experiment 3 is to determine the extent to which dynamic noise in spatial-frequency band j interferes with dynamic ASL signals in band i, for all 16 combinations of $i, j = 1, 2, 3, 4$. Basically, this requires determining the performance versus the signal-to-noise ratio in each of the 16 different band combinations. Because at least half a dozen values of s/n must be sampled to determine a performance function, this experiment requires determination of the performance in almost 100 conditions. Since it is impractical to create and maintain a stimulus set of ASL signs large enough for this immense task, a rating procedure was used instead that involved intelligibility judgments of only two representative ASL signs.

Method

Stimuli

The signals were the recorded ASL signs "home" and "flower" from the previously described set. They were filtered in each of the four bands determined by filter set 3 of experiment 1 (Fig. 1). To generate noise stimuli, we started with white Gaussian noise in (x, y, t). In the frequency domain, the noise power spectrum was shaped, separately in each of the four bands, to conform to the three-dimensional (x, y, t) power spectrum of the signals, that is, within each spatial frequency band, the temporal shape of the noise power spectrum was matched to the shape of the signal temporal spectrum as determined in experiment 2.

Signal Power in a Frame

The signal power in a frame is defined as the variance of the signal luminance over the pixels of that frame. The signal power σ_s^2 is the average power of the frames in a sequence. (In fact, the power variation between frames is small.) The noise power σ_n^2 is computed similarly.

Signal-to-Noise Ratio

The signal-to-noise ratio s/n is σ_s/σ_n. Note that here the signal-to-noise ratio is defined in terms of standard deviations, the root-mean-square (rms) amplitudes of the signal and the noise. These are the square roots of the powers of the signal and the noise. A set of stimuli illustrating the noise, the signals, and their combinations is shown in Fig. 5.

Procedure

The display viewed by the subject consisted of two adjacent sequences. On the left-hand side was a noiseless sign in band i, and on the right-hand side the same ASL sign filtered in the same band i was combined with added noise from band j; 176 such pairs were presented to the subjects. The combinations of i, j, s/n, and the ASL sign occurred in random order.

Rating Scale

Subjects viewed the noisy and noiseless sequences side by side and were asked to rate the noisy one on the following rating scale:

0, Cannot detect sign at all;
1, Barely visible signer, but cannot see sign;
2, Visible signer, some trace of sign;
3, Can guess at sign, but most features indiscriminable;
4, Fairly discriminable sign, but some critical features missing;
5, Visible sign, but poor-quality image;
6, Highly discriminable sign with good-quality image.

Subjects used fractional ratings to describe their judgments more precisely. The noiseless sequences served as references to help the subjects anchor their responses. Ratings were collected from three subjects. Subsequently, the s/n values were adjusted to obtain a better sample of the rating function, and three more subjects were run. In this experiment alone, the subjects were hearing nonsigners.

Results

The stimulus range was quite large, from stimuli in which the subtle details of an ASL sign were perfectly visible to stimuli in which even the presence of the signer was completely masked by noise. Thus the range of ratings, for any particular stimulus condition, was rather small. Within this range, it was most practical simply to treat the ratings numerically and to obtain the average rating across subjects. In a previous study,[9] quality ratings were obtained for a large set of stimuli, a subset of which was then carefully tested by formal intelligibility tests. The correlation between rated quality and objectively measured intelligibility was 0.85. Considering that the intelligibility-tested stimuli were a homogeneous subset of the most-intelligible stimuli, the high correlation was, in the authors' words, "an impressive vindication of the rating procedure" (Ref. 9, p. 364).

Figure 6 shows an example of 1 of the 16 rating-versus-s/n functions for stimulus band 3 with noise band 3. The data (mean rating R versus $\log s/n$) were fitted by three-segment linear functions (a total of three parameters) constrained as follows. (s and n are shown as S and N in all the figures.)

In segment 1, the left-hand asymptote was constrained to be horizontal at $R = 0$. In segment 3, the right-hand asymptote as $s/n \to \infty$ was horizontal at $R = R_*$. Segment 2 connected segments 1 and 3. The square deviation of the data from the three-segment fit was minimized by an optimization program.[19] Figure 6 illustrates the parameter-

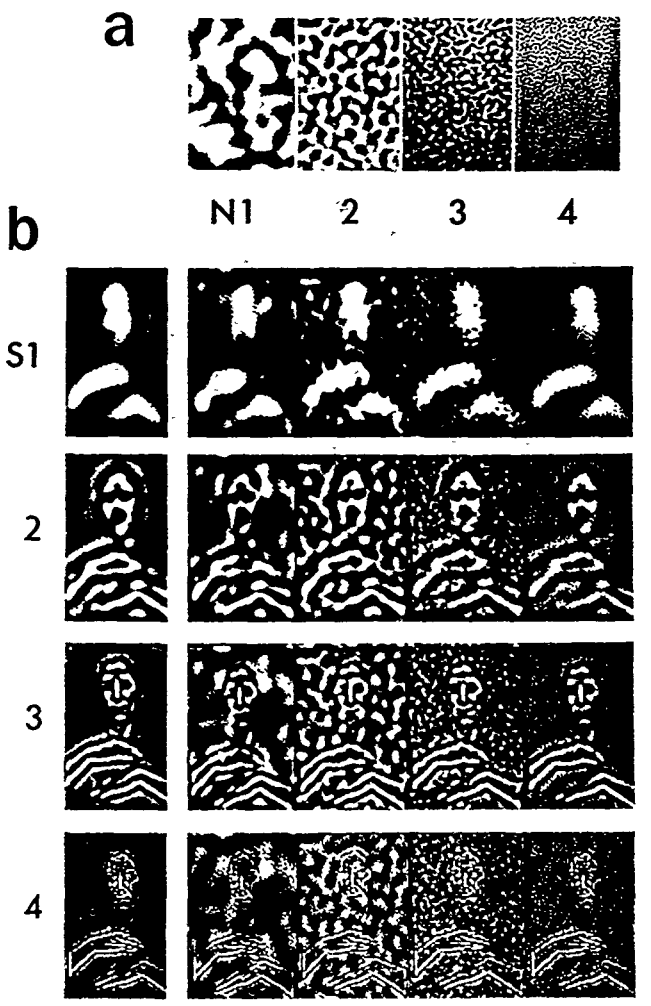


Fig 5 Examples of all combinations of band filtered signals plus band filtered noise a, Gaussian noise filtered in bands 1–4 (left to right) b, Band-filtered ASL signals plus band filtered noise Each row represents a single signal band with band 1 at the top and band 4 on the bottom. Each column (continuing downward from a) represents a single band of Gaussian noise The leftmost column represents the noise free signal.

estimation procedure. The single masking effectiveness parameter $(s/n)_{50\%}$ used to describe each rating function is the s/n ratio at which the function attains 0 5 times its asymptotic height $R\infty$.

Figure 7 shows the set of 16 estimated rating functions that describe the masking of each ASL band by each of the noise bands. The $(s/n)_{50\%}$ values derived from the rating functions of Fig. 7 are graphically displayed in Fig. 8, which summarizes the cross-band-masking data. Bands 1, 2, and 4 mask themselves better than they mask any other band Band 3 appears to mask band 4 slightly more than it masks itself, but we do not have a test of statistical significance for this effect.

Masking as a Function of the Frequency Difference between the Test Stimulus and the Noise Masking Stimulus

Band 1 is more sensitive to masking by noise in its own band than are frequency bands 2, 3, and 4, which, when masking

themselves, are all equally effective; that is, let $(s_j/n_j)_{50\%}$ represent the masking effectiveness of noise band j on signal band i. The points $(s_j/n_j)_{50\%}$, $i = 2, 3, 4$, are all at the same level in Fig. 8; the points $(s_1/n_1)_{50\%}$ is much higher.

To compare band 1 with the other bands, it is necessary to normalize the masking vulnerability of different bands. Masking vulnerability is indexed by self-masking $(s_i/n_i)_{50\%}$. The normalized masking effectiveness NME is

$$\text{NME}\,(s_j/n_j) = (s_j/n_j)_{50\%}/(s_i/n_i)_{50\%}.$$

Masking as a function of the frequency separation between test and noise bands is illustrated in Fig. 9. The abscissa is the ratio f_s/f_n (on a log scale), where f represents the mean frequency of a band. The ordinate represents the log of the normalized masking effectiveness. The straight lines represent a mirror-symmetric function fitted to the

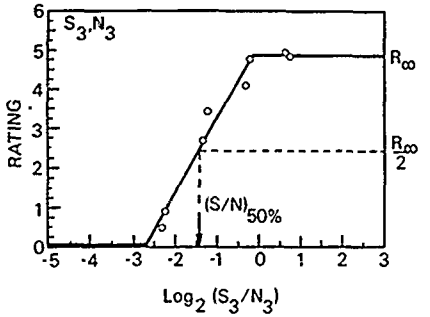


Fig. 6. Average ratings as a function of signal-to-noise ratio for the signal and the noise in band 3. The data are indicated by circles, the three-segment fit is indicated by the heavy lines. The dashed lines indicate the procedure for estimating $(s/n)_{50\%}$, the abscissa value under the arrow

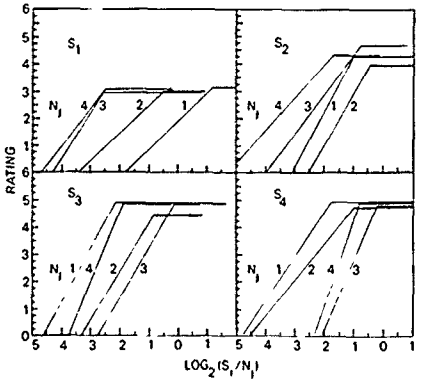


Fig 7 Rating functions for cross band masking The abscissa is the signal-to noise ratio; the ordinate is the mean rating, and the curves represent the three segment best fits to the data Each panel represents data from one signal band s_i, the curve label indicates the band of the noise n_j

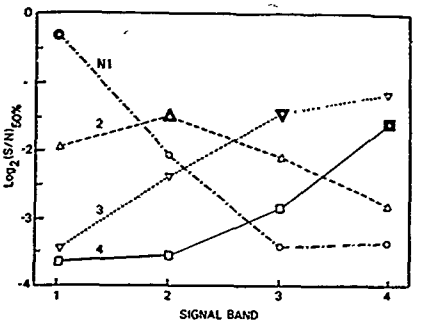


Fig. 8. Masking effectiveness of noise bands against signal bands. The abscissa is the signal band s_i; the ordinate is the value of $(s/n)_{50}$ derived from the rating functions (Fig. 7) by the estimation procedure shown in Fig. 6. The curve parameter indicates the noise band. Emphasized points indicate that the signal and the noise are in the same band

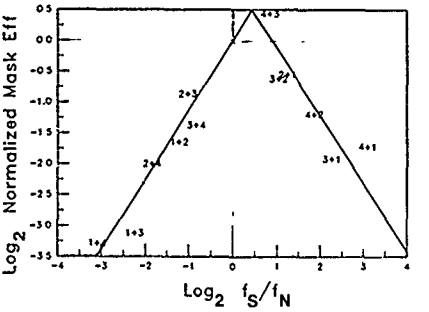


Fig 9. Normalized cross-band masking as a function of frequency separation. Each band is represented by its mean frequency f The abscissa represents the $\log_2$ of f_{signal}/f_{noise} The ordinate is the $\log_2$ of the normalized masking effectiveness, the same data as in Fig 8 with the curves for each signal band i moved up so that $(s_i/n_i)_{50\%}$ falls at 0 0 Signal bands i and noise bands j are indicated by $i + j$, the center of the + indicates the plotted datum The straight lines represent the optimal mirror-symmetric fit to the data, the lines are centered above $\log_2(f_s/f_n) = 0.46$ and with a slope of ± 1.11.

data and constrained to pass through 0, 0. (The mirror-symmetric fit is the most convenient for determining whether there is any asymmetry between the masking effectivenesses of low and high frequencies.) The peak is located to the right of zero; the point of symmetry is $x = \log_2(f_s/f_n) = 0.46$, which represents a frequency ratio for optimal masking of 1:1.38. The slopes of the distance function are ± 1.11.

Cross-band masking is quite adequately described in terms of log frequency separation ($\log f_s - \log f_n$) without the necessity of referencing the particular frequencies that contribute to the separation Masking falls off by a factor of slightly more than 2 when the frequency separation is doubled, a result that is generally consistent with data obtained with much simpler stimuli.[12,13,15] The right-of-center peak in Fig. 9 indicates that noise frequencies lower than the signal mask it slightly better than do frequencies higher than

the signal. This asymmetry is reflected in all six direct comparisons of masking of signal band i by noise band j compared with masking of signal band j and noise band i. For $i > j$, the masking effectiveness $NME(s_i/n_j) > NME(s_j/n_i)$. This masking asymmetry is opposite that obtained with data from simpler stimuli.[20,21]

Although masking falls off with increasing frequency distance between bands, with sufficient power, any noise band can obliterate any signal band; that is, in Fig. 7 all the rating functions were driven to zero at low signal-to-noise ratios. Our spatial-frequency filters are sufficiently narrow that this effect cannot be attributed to common-frequency masking, which occurs when frequencies in the tail of the noise happen to fall within the signal band and are so highly amplified that they change the signal-to-noise ratio within the signal band itself. Most masking between widely separated frequencies is caused by nonlinear distortion in the display system and the visual system, neither of which faithfully reproduces small-amplitude variations in large signals. Both systems, in effect, create masking noise at new frequencies when confronted with high-amplitude inputs. Indeed, the two extreme-left-hand and two extreme-right-hand points in Fig. 9 are at the intensity resolution limit of the display system and might have shown less masking effect (been lower in the figure) had the display system been better able to render small signal-to-noise ratios faithfully. To determine whether masking between widely separated frequencies also arises from genuine channel interactions would require bigger interactions than those observed here. All in all, the cross-band-masking data obtained with our complex displays are quite comparable with data obtained with sinusoidal gratings.

EXPERIMENT 4: ADDING SIGNALS FROM DIFFERENT BANDS

Typically, signal addition has been studied with simple, static signals at low contrast levels in which internal noise is dominant[1,2,22,23] rather than with realistic dynamic stimuli at high contrast levels with high levels of external noise. The purpose of experiment 4 is to discover quantitatively how ASL intelligibility is affected when two dynamic signals from different spatial-frequency bands are algebraically added. The effect on performance of adding two ASL signals is an inherently complex matter because it depends on the signal-to-noise level at which the addition is tested. This dependence is derived in part from the psychometric function (performance versus s/n), which is concave up at low intensities and concave down at high intensities, and in part from more-complex factors. Thus, at high levels of s/n, performance cannot be improved by further increases in s. Insofar as we wish to characterize the efficiency of a detector in terms of internal noise, this would mean that at high input levels, internal noise is proportional to the input.[24]

At low levels of s, performance in detection tasks typically increases with the square of s; i.e., power-law detection is obtained.[25-27] Square-law detection is consistent with constant internal noise, independent of s. Insofar as the square law also applies to band-limited ASL, doubling the amplitude of a signal in band i (and thereby quadrupling its power) might be expected to improve intelligibility more than would adding signal in band j (which would only double signal power).

In contrast, consider the addition of two signals at a high level of s/n. Within any single spatial-frequency band i, even with noiseless stimuli, performance is not so good as in the original unfiltered source images. Therefore, at a high signal level in band i, adding signals from another band j is more effective in improving performance than adding still more signal in band i. Thus different factors are critical for high-intensity and for low-intensity signal combinations, and their combinatorial effects are modeled by different rules.

To study how weak signals combine, we need a method of generating approximately equivalent weak signals. Weakening a signal by reducing the signal contrast relies on the observer's internal noise to weaken the signal. Adding external noise[28] is obviously the better way to control signal intelligibility. Pavel *et al.*[24] showed that for constant s/n, the signal contrast could be varied over a wide range without affecting intelligibility. Indeed, in a preliminary study (see Ref. 10, Exp. 4), this result was verified again with the current set of ASL stimuli. Thus, to study how signals combine, we may use any signals that fall within the enormous range of contrasts that is sufficient to overcome internal noise, and we vary intelligibility by varying external added noise.

Method

Overview

The first step in the procedure is to compose the spatial-frequency amplitude spectrum of an external noise stimulus so that it would mask all signal bands equally. Unfortunately, the rating functions in Fig. 7 are not parallel in the different signal bands, so equal masking of all spatial bands at different intensities is impossible with a single noise source. Given that limitation, we selected a particular noise stimulus to test, first, the intelligibility of weak signals in all bands i under this noise and, second, the intelligibility of all combinations of signals in band i with signals in band j.

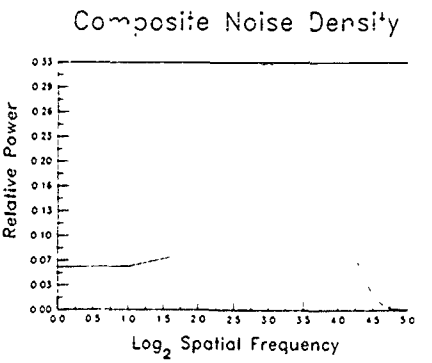


Fig. 10. Spatial power spectrum of the composite noise used in experiment 4. The abscissa is the $\log_2$ of the spatial frequency in cycles per picture width (f_0, the width, is 64 pixels). The extreme left hand side represents 1 cycle per picture, the extreme right hand side represents 32 cycles per picture. The ordinate represents relative power on a linear scale.

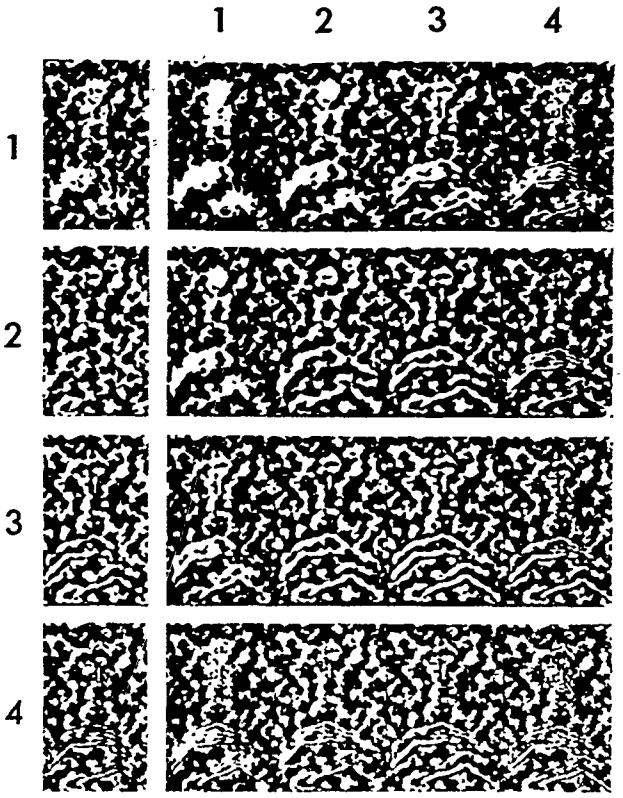


Fig 11 Single frames illustrating the stimuli for experiment 4. The sum of weak signals in bands i and j plus the composite noise of Fig 10 Composite noise is equally present in all stimuli. The leftmost column represents single-band signals, with the band indicated by the number at the left The other panels represent stimuli composed of two signal bands, one component band indicated by the number at the left of the row and the other band indicated by the number at the top of the column.

Composite Noise

From the cross-band-masking data of experiment 3, we inferred a particular composite noise that would be expected to reduce weak signals in all bands to approximately equal intelligibilities. (We use the term composite noise to emphasize that the noise can be regarded as being composed of many spatial-frequency bands, each with a different amplitude and with a different temporal-frequency spectrum.) Full equality of intelligibility may be impossible with any composite noise because of the complex cross-band masking revealed in experiment 3. Figure 10 shows the spectrum of the noise that was used

Signals

The signals were 80 ASL signs, basically the same set that was used in experiment 1b. They were produced at $s_i/n = 0.25$, where s_i indicates the amplitude of signal in band i and n indicates the rms amplitude of the composite noise stimulus. All six combinations of signal in band i with signal band j, $j \neq i$, were produced There were four combinations of signal in band i with itself (i e , $s/n = 0.5$) and four stimuli with signal in band i alone ($s/n = 0.25$) Additionally, a composite signal was composed of the sum of all four bands represented by their amplitude in the $s/n = 0.25$ condition The composite signal was tested alone (the control condition) and under the composite noise (equivalent to $s/n = 1.0$) The stimulus conditions are illustrated in Fig 11

Procedure

The 80 signs were divided to 16 blocks of 5 signs, balanced for difficulty. A Greco-Latin square design was used to generate a completely counterbalanced design in which every block of ASL signs occurred in every signal condition, and the order of conditions was balanced over subjects This required generating 16 different hour long stimulus tapes, one for each of the 16 subjects run in this experiment

The viewing and testing conditions were similar to those described for experiment 1 and particularly for experiment 1b. Subjects were fluent ASL signers from the community. As before, all subjects had good vision under the experimental conditions as determined by an acuity test administered before the experiment.

Results

Figure 12 shows the results for all classes of signals confined to a single band. At $s/n = 0.25$, intelligibility in all bands is below 9%. At $s/n = 0.5$, intelligibility in bands 1 and 2 is 17.5%, whereas performance in bands 3 and 4 is 60.9%. At $s/n = \infty$, the conditions run to test the filters in experiment 1b, intelligibility rises to 66.4% in the lowest band and up to 87.5% in band 3.

Figure 13 shows the same data as Fig. 12 plus the six additional summation conditions of band i with band j, $i \neq j$. The points indicated with circles in Fig. 13 are precisely the same $s/n = 0.5$ points as in Fig. 12. Since they do not seem to fall any differently on the curves than do nearby points that represent different bands, it appears that summation is quite similar within and between bands.

Statistical Analysis of the Data

The design of experiment 4 involves three factors: 16 conditions × 16 subjects × 16 stimulus sets. Because each subject saw each stimulus set only once (and not once in each condition), only 256 of the 4096 possible conditions were run. Typical analysis-of-variance designs are inappropriate for such a sparse design, so a simple linear model was developed. A subject's score y for a set of five stimulus items that constitute a condition ranges from 0 to 5 and is assumed to be the sum of five terms: the grand mean m, factors for condition difficulty c_i, the subject's skill s_j, the ASL set difficulty a_k, and finally a term representing random error $\epsilon_{i,j,k}$:

$$y_{i,j,k} = m + c_i + s_j + c_k + \epsilon_{i,j,k}.$$

Condition difficulty c_i is estimated by

$$c_i = \frac{1}{16}\sum_{j=1}^{16} y_{i,j,k} - m = \frac{1}{16}\sum_{k=1}^{16} y_{i,j,k} - m,$$

that is, by averaging over all subjects and stimulus sets in which condition i occurred and subtracting m. Factors s_j and a_k are estimated similarly. The variance σ^2 of the random error ϵ is $(1/210)\sum\epsilon^2$, where 210 represents the degrees of freedom, the number of cells (256) reduced by the number of estimated parameters (1 + 15 + 15 + 15).

The rms error σ was found to be 0.984. This is approximately what would be predicted from the binomial variability of the data if the predictions $\hat{y}_{i,j,k} = c_i + s_j + a_k$ were based on a completely correct model. The standard error of the mean of the scores shown in Figs. 12 and 13 is ±4.92%.

Summation as a Function of Frequency Distance between Bands

The amount of intelligibility summation as a function of the frequency separation between component signals can be

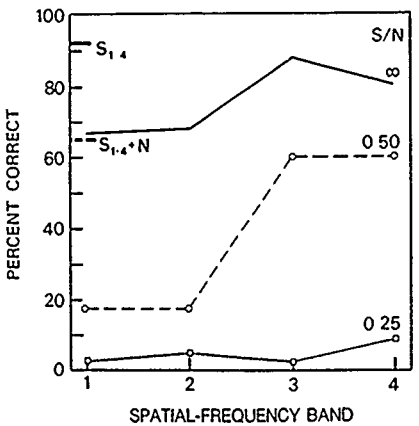


Fig 12. Data from experiment 4. Intelligibility of band-limited single-band signals in composite noise. The abscissa indicates the band of the signal, the ordinate indicates the percent correct scored by the 16 subjects in the intelligibility test. The curve parameter indicates the signal to-noise ratio of the stimuli. The curve labeled ∞ represents data obtained without added noise in experiment 1 (with different subjects and a slightly different stimulus set). On the left-hand ordinate, the point $S_{1\text{-}4}$ indicates intelligibility of the noise-free sum signal of band 1 + band 2 + band 3 + band 4, the point $S_{1\text{-}4}$+N indicates the intelligibility of the same signal plus noise ($s/n = 1$).

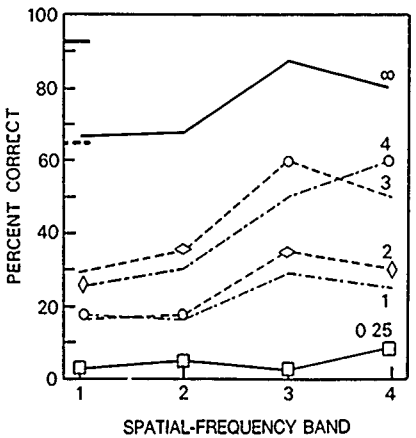


Fig 13. Data from experiment 4. Intelligibility of pairs of band-limited signals in composite noise. The ordinate, the abscissa, and the curves labeled 0.25 and ∞ are as in Fig. 12. The dashed curves indicate signals composed of band i (indicated on abscissa) and band j (indicated as the curve parameter). The open circles represent data for $i = j$, the middle curve of Fig. 12. The flat diamonds represent the addition of nearby signal bands (2 and 3), the tall diamonds represent the addition of distant bands (1 and 4). The pairs indicated by diamonds are matched for the strengths of their constituent signals.

tested nicely by using the data of experiment 4. Because, at $s/n = 0.5$, bands 1 and 2 have, by coincidence, exactly the same intelligibility (17.5%) and bands 3 and 4 have the same intelligibility (60.0%), we compare the intelligibility of band 1 plus band 4 (wide separation) with that of band 2 plus band 3 (small separation). These two points are at slightly different intelligibility levels in Fig. 13; the small band separation (flat diamonds) at 35% is somewhat higher than the large separation (thin diamonds) at 25%. The probability that a difference this large would occur by chance, estimated by a one-tail z test, is 0.040.

To determine whether it is more efficient to improve a weak signal in band i by adding more energy in i or do so by adding energy in an adjacent band j, we compare the effects of summing two signals at $s/n = 0.25$. In Fig. 13, the crossings of the curves labeled 3 and 4 at the extreme right and the crossings of the curves labeled 1 and 2 at the extreme left indicate that there is a tendency for the sum of band 4 + band 4 ($I = 60\%$) and of band 3 + band 3 ($I = 60\%$) to be more intelligible than band 3 + band 4 ($I = 50\%$) and for the sums band 1 + band 1 and band 2 + band 2 (both $I = 17.5\%$) to be slightly more intelligible than band 1 + band 2 ($I = 16.3\%$). The probabilities of these differences' occurring under the null hypothesis are 0.024 and 0.209, respectively. Taken together, these observations imply that, with the signal levels studied here, there is a small but occasionaly significant tendency for component signals to contribute more to intelligibility when they are closer in frequency.

Efficiency When Signal Power Is Constrained

For practical purposes, when two different weak visual ASL signals are summed, the effect of frequency separation on intelligibility is small. All the factors that might have contributed to a separation effect or an inverse separation effect are almost in balance at the s/n values investigated here. To improve intelligibility, given a signal in band i, adding more signal in any other band j is almost as effective as adding more signal in i. In these signal manipulations, we are speaking of signal amplitudes. If we were concerned with signal power rather than with rms amplitude, then it would clearly be more efficient to distribute the power over different bands. Doubling the amplitude within a band quadruples the power, whereas the power of signals in disjoint bands adds linearly.

SUMMARY AND CONCLUSIONS

(1) In low-resolution dynamic ASL images (96×64 pixels), it is possible to divide the original signal into four different frequency bands, each of which is quite intelligible (67–87% for isolated ASL signs) and each of which could serve for ordinary ASL communication.

(2) The empirically determined temporal-frequency spectrum of ASL is approximately the same in all spatial-frequency bands.

(3) The ratio of root-mean-square signal amplitude to noise amplitude, s/n, at which ASL becomes intelligible is nearly the same for the three highest bands, but the critical s/n is higher for the lowest-frequency band.

(4) The masking of signals in one band by noise in another is governed simply by the ratio of frequencies between the bands (the difference of the log frequencies). There is asymmetry: noise lower in spatial frequency than the signal is more effective in masking than is higher-frequency spatial noise. When the frequency separation between signal and noise is increased by a factor of 2, intelligibility can be maintained at 1/2 the original signal-to-noise ratio.

(5) When two weak signals ($s/n = 0.25$) are added, the intelligibility of the summed signal is slightly greater when the two signals are in adjacent frequency bands than when they are widely separated bands; and intelligibility is slightly greater when the two signals are identical than when they are in adjacent bands. If the signal power—not amplitude—is limited, intelligibility is maximized by dispersing the signal power widely across frequency bands.

APPENDIX A: FILTER-GENERATION ALGORITHM

This algorithm generates K filters that divide frequency space (ω_x and ω_y) into partially overlapping annular regions whose boundaries are adjustable. The summed output of all the filters equals the original input signal.

Let K be the desired number of filters. Let LP represent the Fourier transform of a low-pass filter; that is, $||\,\mathrm{LP}(\omega_x, \omega_y)\,||$ is monotonically decreasing in ω_x and ω_y. (The particular LP_i that are used to feed the algorithm are defined below.) We use the terms center and surround analogously to their use in composing difference-of-Gaussian filters; they refer to x, y spread functions of the filters. The center and surround components are used as kernels to generate the filters. The surround of filter $K - i + 1$ becomes the center of filter $K - i$ (the next lower filter in terms of frequency). In the sum of all the filters, all the centers and surrounds cancel, and the original source image is recovered. The steps in the algorithm are stated in terms of the two-dimensional Fourier transforms of the filters and their components:

(1) Define F_K, the highest-frequency filter. The center of F_K is defined to be $C_K = 1$. The surround of F_K is defined in terms of LP_K (see below) as $S_K = 1 - (1 - \mathrm{LP}_K)^m$, then the Kth filter is $F_K = C_K - S_K = (1 - \mathrm{LP}_K)^m$.

(2) Do the following loop $K - 2$ times ($i = 1, K - 2$) to generate, in sequence, the filters $K - 1, K - 2, \ldots, 2$.

(a) Define the center of the $K - i$ filter as the surround of the previously defined filter: $C_{K-i} = S_{K+1-i}$.

(b) Define the surround of the $K - i$ filter. $S_{K-i} = 1 - (1 - \mathrm{LP}_{K-i})^m$. The surround is a low-pass filter derived from a generating low-pass filter LP_{K-i} chosen so that S_{K-i} will have a lower cutoff frequency than C_i in accordance with the desired partition of frequency space.

(c) Define the $K - i$ filter as the center minus the surround: $F_{K-i} = C_{K-i} - S_{K-i}$.

(d) Increase i, if $i \leq K - 2$, return to step (a), otherwise, continue to step (3).

(3) $F_1 = 1 - \sum_2^K F_i = S_2$, that is, F_1 is the low-pass filter that was chosen as the surround of F_2, it encompasses all the residual signal. Note that $\sum_1^K F_i = 1$.

To begin the algorithm with $F_K = (1 - \mathrm{LP}_K)^m$, LP_K must be defined. Let LP_K be a two-dimensional Gaussian low-pass filter whose frequency-domain representation is

$$LP_K(\omega_x, \omega_y, \sigma_x, \sigma_y) = \exp[-2\pi^2(\sigma_x^2\omega_x^2 + \sigma_y^2\omega_y^2)],$$

where ω_x and ω_y are the frequency components in the x and y directions, respectively, and σ_x and σ_y are the x and y widths of the generating spatial Gaussians. Since $F_K = (1 - F_K)^m$, as m increases, the frequency cutoffs become steeper, and the overlap between filters is reduced (which is good); but for $m > 4$, the ringing in the x-y-space domain becomes obtrusive (which is bad). Therefore $m = 4$ was chosen.

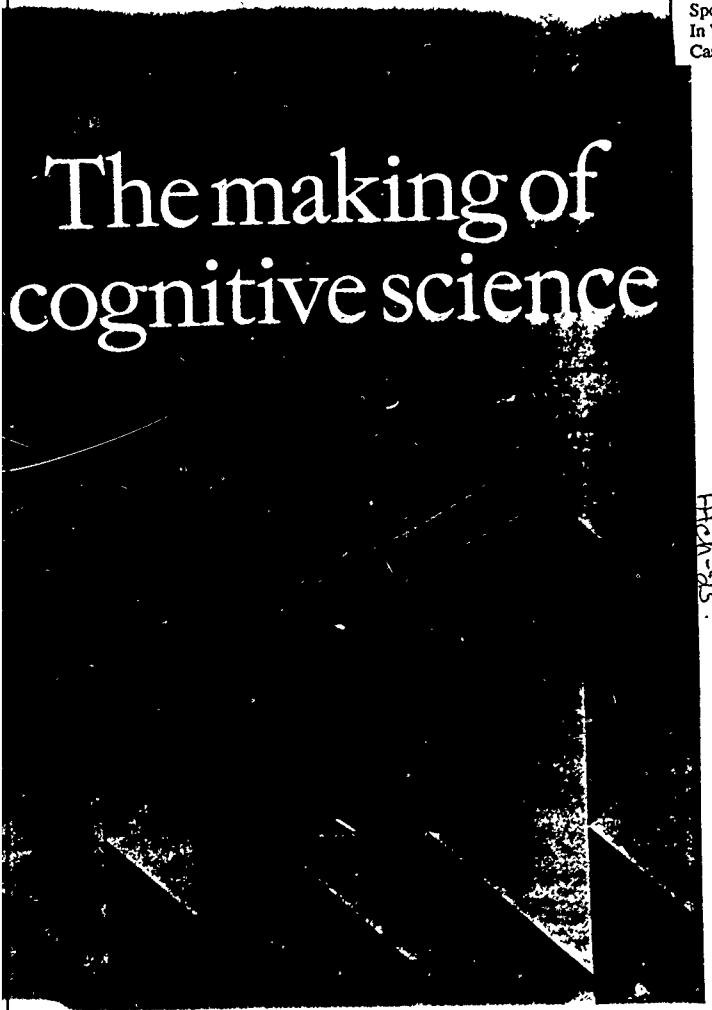
ACKNOWLEDGMENTS

This research was supported partly by National Science Foundation, Science and Technology to Aid the Handicapped, grant no. PFR-80171189, and by U.S. Air Force Life Sciences Directorate grant AFOSR-85-0364. The work was conducted at New York University and partially fulfilled the requirements for a Ph D. degree in experimental psychology for Thomas R. Riedl. Throughout the work, we received valuable advice and guidance from Misha Pavel and Michael S. Landy. We thank James P. Thomas for his comments on the manuscript and Geoffrey Iverson for statistical advice. We express our gratitude for the help that we have received from many individuals in the deaf community and from organizations, including The Disabled Student Services Office of New York University, The New York Society for the Deaf, and D E.A.F. We thank our signer, Ellen Roth, and we wish to acknowledge the skillful technical assistance of Robert Picardi and August Vanderbeek.

* Present address, AT&T Bell Laboratories, Whippany Hill, New Jersey 07981, requests for reprints should be addressed here

REFERENCES

1 For a succinct review, see N Graham, "Detection and identification of near threshold visual patterns," J Opt Soc Am A 2, 1468–1482 (1985)
2 For a review, see L A Olzak and J P Thomas, "Seeing spatial patterns," in *Handbook of Perception and Performance*, K R. Boff, L Kaufman, and J P Thomas, eds (Wiley, New York, 1986), Chap 7
3 G. E Legge, D G Pelli, G S Rubin, M M Schleske, "Psychophysics of reading—I. Normal vision," Vision Res 25, 239–252 (1985)
4 D. H Parish and G Sperling, "Object spatial frequency, not retinal spatial frequency, determines identification efficiency," Invest Ophthalmol Vis Sci Suppl 28, 359 (1987)
5 G Sperling and D. H. Parish, "Object spatial frequencies, retinal spatial frequencies, noise, and the efficiency of discrimination," in *Mathematical Studies in Perception and Cognition* (Department of Psychology, New York University, New York, N.Y., 1987).
6 J D Schein and M T Delk, Jr , *The Deaf Population of the United States* (National Association of the Deaf, Silver Spring, Md , 1974).
7. G. Sperling, "Bandwidth requirements for video transmission of American Sign Language and finger spelling," Science 210, 797–799 (1980).
8. G. Sperling, "Video transmission of American Sign Language and finger spelling: present and projected bandwidth requirements," IEEE Trans. Commun. COM-29, 1993–2002 (1981)
9. G. Sperling, M. Landy, Y. Cohen, and M. Pavel, "Intelligible encoding of ASL image sequences at extremely low information rates," Comput. Vision Graph. Image Process. 31, 335–391 (1985).
10. T. R. Riedl, "Spatial frequency selectivity and higher level human information processing," doctoral dissertation (New York University, New York, N.Y., 1985).
11. P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," IEEE Trans. Commun. COM-31, 532–540 (1983).
12. G. B. Henning, B. G. Hertz, and J. L. Hinton, "Effects of different hypothetical detection mechanisms on the shape of spatial-frequency filters inferred from making experiments. I Noise masks," J. Opt. Soc. Am. 71, 574–581 (1981).
13. G. Legge and J. Foley, "Contrast masking in human vision," J Opt. Soc. Am 70, 1458–1471 (1980).
14. S. Stecher, C. Sigel, and R. V. Lange, "Composite adaptation and spatial frequency interactions," Vision Res. 13, 2527–2531 (1973).
15 C F. Strohmeyer III and B Julesz, "Spatial frequency masking in vision: critical bands and spread of masking," J. Opt. Soc Am. 62, 1221–1232 (1972).
16 C. F. Strohmeyer III, S. Klein, B M Dawson, and L Spillmann, "Low spatial-frequency channels in human vision adaption and masking," Vision Res. 22, 225–233 (1982)
17. D. J. Tolhurst, "Adaptation to square-wave gratings. Inhibition between spatial frequency channels in the human visual system," J. Physiol. 226, 231–248 (1972)
18 H. R. Wilson and J. R Bergen, "A four mechanism model for threshold spatial vision," Vision Res 19, 19–32 (1979).
19. J P. Chandler, "STEPIT," in *Quantum Chemistry Program Exchange* (Department of Chemistry, Indiana University, Bloomington, Ind., 1965).
20 J Nachmias and A. Wever, "Discrimination of simple and complex gratings," Vision Res. 15, 217–223 (1975)
21 D J Tolhurst and L P. Barfield, "Interactions between spatial frequency channels," Vision Res 18, 951–958 (1978)
22 N. Graham and J. Nachmias, "Detection of grating patterns containing two spatial frequencies a comparison of single-channel and multiple models," Vision Res. 11, 251–259 (1971)
23 N. Graham, "Psychophysics of spatial frequency channels," in *Perceptual Organization*, M. Kubovy and J Pomerantz, eds (Erlbaum Halstead, Potomac, Md., 1980), pp 1–25
24. M. Pavel, G Sperling, T. Riedl, and A. Vanderbeek, "Limits of visual communication: the effects of signal to noise ratio on the intelligibility of American Sign Language," J Opt Soc Am A 4, 2355–2365 (1987).
25 C R Carlson and R. W Klopfenstein, "Spatial frequency model for hyperacuity," J Opt Soc Am A 2, 1747–1751 (1985)
26 J Nachmias and R. V. Sansbury, "Grating contrast discrimination may be better than detection," Vision Res 14, 1039–1042 (1974)
27 C F. Strohmeyer and S Klein, "Evidence against narrow-band spatial frequency channels in human vision the detectability of frequency modulated gratings," Vision Res 15, 899–910 (1975)
28 D. G. Pelli, "Effects of visual noise," doctoral dissertation (University of Cambridge, Cambridge, 1981).



Sperling, G. The magical number seven: Information processing then and now.
In William Hirst (Ed.), *The making of cognitive science: Essays in honor of George A. Miller*
Cambridge, U.K.: Cambridge University Press, 1988. Pp. 71-80

5 The magical number seven: information processing then and now

George Sperling

The magical hypotheses

George Miller read an invited address to the Eastern Psychological Association on April 15, 1955: "The Magical Number Seven, Plus or Minus Two: Some Limits in Our Capacity for Processing Information." He complained: "I have been persecuted by an integer." The article was published the following year in the *Psychological Review*. Both the spoken version and the written version (here referred to as 7 ± 2) were immediate successes. A survey conducted twenty years later found that 7 ± 2 was the single most often cited paper in cognitive psychology (Garfield, 1975). Every student of cognitive psychology has been exposed to it, and many cognitive psychologists, including myself, have been profoundly influenced by it.

The central thesis of 7 ± 2 is that the number 7 occurs in two contexts. The first context is absolute judgments of brightness, loudness, pitch, extent, and so on. In absolute judgments – that is, classification of sensory stimuli into categories – 7 ± 2 *is* the effective number of categories that the subject can maintain. This number is derived from what was then a novel statistical computation: the 2 to 3 bits of *information transmitted* by an observer in these tasks.

To transmit more than 3 bits of information in an absolute judgment, a subject requires sensory stimuli that vary in more than one dimension. The transmitted information in each component dimension suffers somewhat as new dimensions are added; nevertheless, subjects can transmit enormously greater amounts of information about multidimensional stimuli than about one-dimensional stimuli.

The second context in which 7 ± 2 occurs is memory. In short-term recall (the classical immediate memory test), 7 ± 2 describes the number of items that a subject can recall. In contrast to the absolute judgments, the number of recallable items does not depend on their information content, so the information-transmitted statistic does not predict performance. Although binary digits contain less than one-third the information of decimal digits, a subject can recall only very slightly more binary digits than decimal digits. However, by recoding sev-

eral separate elements (e.g., binary digits) into a unitary "chunk" (e.g., an octal digit), the subject can enormously increase his or her recall capacity for binary digits.

George Miller classified many of the experimental procedures that had been used to study information processing into two categories (absolute judgments and immediate recall) and invented twin hypotheses, one to characterize the human performance limit in each category. One-dimensional absolute judgments could transmit up to 3 bits of information; immediate recall was limited to 7 ± 2 chunks. All of the experiments that George Miller advanced in support of these twin hypotheses had been performed by others. The contribution of 7 ± 2 is entirely theoretical – its succinct classification of a great deal of data and its clear formulation of two hypotheses that demanded theoretical explanations.

Professor George Miller

In the spring of 1958, Professor Miller offered a seminar in information processing for the graduate students in the Psychology Department at Harvard. I enrolled in that class and volunteered to give the seminar presentation on 7 ± 2. How I got to this point is a story in itself.

The previous spring, I had been a student in the neighboring Department of Social Relations. Professor Miller conducted the last sessions of the required proseminar (offered jointly with Jerome Bruner, Richard Solomon, and George Mandler). I presented a class report on a paper by Lawrence and Laberge (1956). After describing their experiments, I proposed a partial report experiment that could better address the same issues. Professor Miller was sufficiently interested in this proposed experiment to offer to support it. Thereby began our association.

In the summer of 1957, Professor Miller obtained permission for me to use Jerry Bruner's tachistoscope during Bruner's absence, and he supported my application for a transfer to the Psychology Department. In the fall, just as I entered the Psychology Department, my draft board officially notified me that they now expected me to turn my attention to their needs. This was obviously to be my last chance in graduate school. Fortunately, during the summer, with the encouragement of Roger Shepard (Professor Miller's postdoctoral fellow, whom he designated to oversee the research), I had completed the experiments for my Ph.D. thesis on what Ulric Neisser (1967) later dubbed "iconic memory."

In those days, from the students' point of view, Harvard's Psychology Department was clearly divided on ideological grounds that were reflected in its geographical layout. Fred Skinner was located at the north end of the basement of Memorial Hall, Smitty Stevens was entrenched at the opposite end, and George Miller, Phil Teitelbaum, Edwin Newman, and everyone else who had not chosen

sides was thrown together in the middle. At the opposite ends of Memorial Hall, students attended competing discussion groups offered weekly by Stevens and Skinner. Cognitive science was centered in the middle.

In Professor Miller's seminar, each student was to present an extended two-hour seminar on an important paper. I chose 7 ± 2 because it was supposed to be a very important paper. On first reading, 7 ± 2 seemed to offer the student an opportunity for a dazzling display of critical acumen. Its assertion that the span of absolute judgment could be so low was patently absurd. Simply searching out the references would uncover their procedural artifacts. For visual judgments of spatial extent, this nitpicking paid off. The displays were tiny (subtending less than ½ degree) and, even so, 3.25 bits of information were transmitted (corresponding to 9.5 alternatives). I imagined that larger displays would undoubtedly produce still better performance and thereby significant violations of 7 ± 2. In all of the other modalities, however, the data withstood scrutiny, and they firmly contradicted my intuition. I almost began to feel persecuted, too.

Miller had noted that, with multidimensional stimuli in absolute judgment experiments, the additional dimensions enable the subject to surpass the 7 ± 2 restriction on the number of effective categories. I intended to propose using this property in reverse to discover what the underlying dimensions of judgments were. Since this seminar was often attended by postdoctoral fellows (such as Roger Shepard) and students from Social Relations (such as Saul Sternberg), as well as by the Psychology Department students, it was an occasion for lively exchanges. However, on this occasion, another psychology student, Jerry Shickman, objected so vehemently and persistently to my use of the concept of dimension that I was unable to proceed. By the time Professor Miller intervened to get things going again, so much tension had built up that subsequent discussion was totally inhibited.

The seminar meeting was a failure. However, the intellectual seed had been sown. The particular set of problems and the issues surrounding them have remained with me ever since. And although I have been persecuted by editors and critics for many more than seven years, my experiments and the data they generated have been a source of much comfort.

The challenge of a theory

What was it about 7 ± 2 that made it such a milestone in cognitive psychology? As usual, it was not just one thing but a propitious combination of many factors. The framework of the presentation was masterful: the provocative challenge of the theoretical approach. Like Sherlock Holmes, the theoretician demonstrates that the evidence is already at hand. One need only be clever enough to perceive

it. By implication, of course, the data-bound experimenters and the rest of us in the audience were not up to the task. It was unnecessary for George Miller to proclaim ponderously that there must be a role for theoreticians in a discipline dominated by experimentalists; the demonstration was a better proof. A cognitive psychologist could be a theoretical psychologist proposing simple hypotheses that organized large amounts of data.

The classical period and the dark ages of information processing

To understand why George Miller's twin hypotheses – his theory – had such an impact, we must examine the field as it was when 7 ± 2 burst forth upon it. The three main textbooks of the time were Woodworth and Schlosberg's *Experimental Psychology* (1954), Osgood's *Method and Theory of Experimental Psychology* (1953), and Stevens's *Handbook of Experimental Psychology* (1951).

Osgood offers no treatment whatsoever of any of the paradigms, data, or issues raised in 7 ± 2. In Stevens's *Handbook*, C. H. Graham's chapter on visual perception has a subsection suggestively entitled "Span of Perceptions," but its four pages are devoted entirely to Hunter and Sigler's (1940) study of estimated dot numerosity as a function of luminance and exposure duration in brief displays. The chapter on cognitive processes by Leeper is utterly astounding in terms of how the field is defined today. The beginning is bogged down in a discussion of consciousness, and the remainder is devoted to concept formation in monkeys.

Not only George Miller's interests in 7 ± 2 (information transmitted in perceptual judgment and in short-term recall) were given short shrift in Osgood's and Stevens's compendia of the 1950s. Attention, in the context of human performance, is entirely absent. Yet, in classical psychology, attention is the heading under which the paradigms of 7 ± 2 would be treated. Related subjects were also ignored. For example, even a subject with sensory as well as cognitive implications – saccadic eye movements as our means of acquiring visual information – is not mentioned in either text. Anyone who thinks that 7 ± 2 did not represent a leap forward in our conceptualization of the important issues of psychology need only look at the primordial ooze from which it sprang.

Woodworth and Schlosberg's views of the important issues in psychology fare better by today's standards. Most of the topics previously mentioned are considered. The treatment is dust bowl empirical; many experimental procedures are summarized and the results stated matter-of-factly. If one already knows what the interesting questions are, the data speak for themselves. But there is not a trace of theory.

The treatment of spans in Woodworth and Schlosberg derives directly from the earlier edition of Woodworth (1938). And this book, in turn, owes much to Woodworth's long internship in Wundt's laboratory.

The debt to Wundt

Wundt was concerned with precisely the issues that 7 ± 2 raises and devoted considerable attention to them immediately in Chapter 1 of his popular *Introduction to Psychology* (1912), a condensation of his longer *Outlines*. Wundt asks, how many elements can be maintained simultaneously in consciousness? When the elements are ticks of a metronome, the answer depends on how the subject groups them. Subjectively grouping ticks by twos (imagining 2/8 time) yields sixteen conscious clicks grouped into eight "chunks" (to use Miller's term). Grouping by eights (imagining 4/4 time) yields only five chunks, although chunking increases the total number of ticks from sixteen to forty. Similar advantages of chunking appear in the visual domain when viewing clearly visible tachistoscopic flashes of unrelated letters or words. Viewers report they can perceive as many unrelated words as unrelated letters (about four). In the auditory modality, observers can report back six spoken nonsense syllables. Braille characters are designed to use only six tactile dot positions. Basically, Wundt proposes the magical number 6 ± 2, and indeed, for immediate memory, this describes the situation.

Unlike his mentor, Helmholtz, whose procedures are as timely today as they were one hundred years ago, Wundt's usual methods are entirely subjective. This was by Wundt's design. When he thought the occasion required it, Wundt could be quantitative and rigorous by today's standards. For example, his *Physiological Psychology* is full of examples of the comparative structure of sense organs in various species and of mathematical formulations of significant relations. Wundt was versed in biology and mathematics. In trying to forge a distinctive science of psychology, different from physiology, Wundt wanted to use distinctively different procedures. Thereby, he chose the wrong path. The new psychology was not then, and still is not, ready for Wundt's introspective methods. Cognitive psychology still succeeds best with experimental procedures that place the complexity in the stimulus and leave the response simple and constrained. Wundt's reverse procedures, such as presenting the subject with a simple red patch and asking him to introspect at length about what he sees, are still beyond our reach.

Helmholtz and Wundt raised many questions that we regard as core issues of cognitive psychology. Helmholtz concentrated primarily on perception, and his methods were universal. Wundt went further, to information processing, thinking, and beyond, but his methods too often were introspective. Many of Wundt's followers were less well versed in scientific protocol than he. In their hands, the

introspective methods ultimately stimulated psychologists to create the behavioral revolution. Unfortunately, as in many revolutions, the good was overthrown with the bad. For almost half a century, not only the methods but even the questions were discarded by most American psychologists.

Traces of interest in cognition survived in a new empirical garb, as noted, for example, by Woodworth and Schlosberg. Although casting cognitive questions in empirical, behavioral terms was an improvement, in the absence of theory the spark was lost. Aside from observing empirical relationships, as in the span of apprehension experiments, there was no inkling of how to cast theories.

The renaissance

When Shannon's (Shannon & Weaver, 1949) information theory came on the scene, it was quickly adapted to a variety of paradigms because it was the only systematic framework psychologists had for dealing with information. [At the same time, Wald's (1950) statistical decision theory blossomed in psychology as signal detection theory.] But the routine application of information theory to psychological paradigms was unfruitful. In the morass of information-oriented experiments, 7 ± 2's contribution was the clear delineation of a domain in which information theory was useful (absolute judgments) and a complementary domain (short-term list recall) in which it was not. This was an important and necessary step in moving forward. In a larger scale, in relation to its time, 7 ± 2 redefined the classical subject of span experiments in terms of information processing, an area of cognitive psychology with clearly phrased problems and the possibility of significant theoretical approaches.

The baroque era and the age of computers

The classical past offered the fascinating cognitive issues posed by Wundt and, in the United States by James but completely lacked an adequate methodology. In the post–World War I period of unmitigated empiricism, the intellectual thread was lost. A post–World War II flurry of information theoretic studies culminated in 7 ± 2. What was and is yet to come?

In outline, the path ahead looks straightforward. The 7 ± 2 theory is a descriptive macro theory. That is, it offers mathematical descriptions of stimulus–response relations at a very global level. The descriptive formulations of 7 ± 2 will eventually be supplemented with process theories – models that embody the step-by-step computations carried out in the cognitive microprocesses that underlie performance. Eventually, the process models will be fleshed out with neural components that represent the biological structures that carry out the cognitive

microprocesses. The early stages of this process of scientific evolution can already be discerned.

Acoustic confusibility

The first important progress following 7 ± 2 came with the observation that not all items were equivalent, even when they conveyed precisely the same information. It was discovered that the number of items recalled from visual or auditory presentations depends on the acoustic structure of the items. Items that are acoustically confusible (such as the letters *b, c, d, g, p, t,* and *v*) are not recalled as well as items that are less confusible (Conrad & Hull, 1964; Sperling, 1963; Sperling and Speelman, 1970). Thus, all chunks are not equivalent; how well a chunk is remembered depends on its sound.

That acoustic structure is critical suggests an acoustic basis for short-term memory (Sperling, 1968). Articulatory coding is an alternative possibility (Hintzman, 1967). There are severe difficulties with any purely structural theory – acoustic or articulatory – since familiarity, which is not easily embodied in any of these theories, has an enormous role in short-term recall (Sperling, Parish, Pavel, & Desaulniers, 1984). Any contemporary theory of short-term recall must deal in much more detail with much more detailed cognitive processes than Miller was forced to do. For example, we know that the phonemic and acoustic structure of the to-be-remembered chunks matters; that recoding, rehearsal strategies, and grouping have specifiable mnemonic effects; and that prior learning experiences with the items are critical. These are some of the presumed components of process theories.

Neural models for immediate memory have been proposed by Grossberg (1980), bypassing the functional process description. However, in the absence of a functional model to explain recoding, rehearsal, grouping, and other strategic options available to the subject, neural specification is probably premature.

Memory noise in absolute judgments

The limiting factor in absolute judgments seems to be that subjects cannot remember the precise boundaries of their categories. There is some uncertainty (noise) in coding stimulus intensity, but the main bottleneck seems to be limited capacity memory (Durlach and Braida, 1969).

One particular formulation of the memory bottleneck (Heinemann, 1984) has been extensively tested on pigeons as well as humans. Chase (1984) and Heinemann find that pigeons make absolute judgments of auditory intensity that are qualitatively quite similar to human judgments. They model the limited capacity memory by assuming that, in memory, the outcome of a trial is represented by a

record containing the stimulus, the response, and the feedback on that trial. Memory contains about 1,000 locations. Each new trial is stored independently at a random location of memory, overwriting the previous contents of the location. When it comes time to make a judgment of a new stimulus, some records, typically estimated to be about seven, are extracted from the 1,000-location long-term memory and placed into a short-term working memory. The judgment is made by comparing the unknown stimulus to the contents of working memory.

Heinemann's model accounts nicely for a number of second-order effects relating to absolute judgments, such as how discrimination in various parts of the stimulus dimension depends on the spacing of stimuli and how a relabeling of the stimuli is gradually learned. Although Heinemann avoids the inference, for me the attractive aspect of this kind of process model is that the working memory that holds the records of previous trials of the judgment experiment may be the same memory that holds the chunks in the immediate memory experiment. This kind of theory is representative of the exciting kinds of models of mental microprocesses that cognitive psychology offers. It also illustrates the incompleteness of theories in which the complex control processes needed to utilize limited-capacity memories are axiomatically assumed rather than explicitly modeled.

The coming era

Physiologically, it is unlikely that formatted records of the sort described in Heinemann's model are written in a memory with a fixed number of locations. This description of a process is best regarded as a convenient, workable conceptualization of a neural network. Indeed, it is easy to design neural networks that behave like the stack memory previously described. Whether the network that actually performs the stacklike memory function in the brain is describable in simple terms is not known. It may or may not be simply organized. However, a complete functional description that can be applied to any particular experimental situation will undoubtedly be very complex.

One of the remarkable emerging properties of the many recently proposed neural networks is that they are quite similar in their overall learning properties, even though their structures are profoundly different. These networks can be used interchangeably to fill the black boxes of the functional process models in much the same way that computer memory chips of different manufacturers can be interchanged on a central processing unit board. At least, that is how one aspect of the future of cognitive models appears to me in the 1980s.

The future

What is the future of 7 ± 2? It offers a descriptive theory of two classes of phenomena. We look forward to better process theories and eventually, to physiologically plausible neural theories. Does this mean that 7 ± 2 will become outmoded and replaced? Not necessarily. Predicting the fate of any psychological theory requires a careful consideration of the more general role of the theory.

The goal of theory in experimental psychology is to provide the best possible description of a class of phenomenon at a given level of complexity (Sperling, 1978). Subsequent theories may be more complex and detailed, but they will not replace earlier theories unless they can explain more with equal or less complexity. Insofar as a theory offers the best description at a given level of complexity, it is eternal and will not be replaced, though it certainly will be supplemented. There are many difficulties with this formulation of the goal of theory, not the least of which is the continuously changing nature of complexity. But the magical cognitive powers of the number seven make 7 ± 2 a probable candidate for the best theory at its chosen level of complexity. If so, 7 ± 2 will endure. That is the ultimate achievement of any theory.

References

Chase, S. 1984. Pigeons and the magical number seven. In M. L. Commons, R. J. Herrnstein, & A. R. Wagner (Eds.), *Quantitative Analyses of Behavior: Discrimination Processes* (pp. 36–47). Cambridge, Mass.: Ballinger.

Conrad, R., & Hull, J. A. (1964). Information, acoustic confusion, and memory span. *British Journal of Psychology, 55*, 429–32.

Durlach, N. I., & Braida, L. D. (1969). Intensity perception. I. Preliminary theory of intensity resolution. *Journal of the Acoustical Society of America, 46*, 372–83.

Garfield, E. (1975). Highly cited articles. 19. Human psychology and behavior. In E. Garfield, *Essays of an Information Scientist* (pp. 262–8). Philadelphia: Institute for Scientific Information.

Graham, C. H. (1951). Visual perception. In S. S. Stevens (Ed.), *Handbook of Experimental Psychology* (pp. 868–920). New York: Wiley.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review, 87*, 1–51.

Heinemann, E. G. (1984). A memory model for decision processes in pigeons. In M. L. Commons, R. J. Herrnstein, & A. R. Wagner (Eds.), *Quantitative Analyses of Behavior: Discrimination Processes* (pp. 3–19). Cambridge, Mass.: Ballinger.

Hintzman, D. L. (1967). Articulatory coding in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 6*, 312–16.

Hunter, W. S., & Sigler, M. (1940). The span of visual discrimination as a function of time and intensity of stimulation. *Journal of Experimental Psychology, 26*, 160–79.

Lawrence, D. H., & Laberge, D. L. (1956). Relationship between accuracy and order of reporting stimulus dimensions. *Journal of Experimental Psychology, 51*, 12–18.

Leeper, R. (1951). Cognitive processes. In S. S. Stevens (Ed.), *Handbook of Experimental Psychology* (pp. 730–57). New York: Wiley.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Neisser, U. (1967). *Cognitive Psychology*. New York: Appleton-Century-Crofts.

Osgood, C. E. (1953). *Method and Theory in Experimental Psychology*. New York: Oxford University Press.

Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Sperling, G. (1963). A model for visual memory tasks. *Human Factors, 5*, 19–31.

Sperling, G. (1968). Phonemic model of short-term auditory memory. *Proceedings, 76th Annual Convention of the American Psychological Association, 3*, 63–4.

Sperling, G. (1978). *The Goal of Theory in Experimental Psychology*. Bell Telephone Laboratories Technical Memorandum 78-1221-12.

Sperling, G., Parish, D. H., Pavel, M., & Desaulniers, D. H. (1984). Auditory list recall: Phonemic structure, acoustic confusibility, and familiarity. *Bulletin of the Psychonomic Society, 18*, 36.

Sperling, G., and Speelman, R. G. (1970). Acoustic similarity and auditory short-term memory: Experiments and a model. In D. A. Norman (Ed.), *Models of Human Memory* (pp. 149–202). New York: Academic Press.

Stevens, S. S. (Ed.). (1951). *Handbook of Experimental Psychology*. New York: Wiley.

Wald, A. (1950). *Statistical Decision Functions*. New York: Wiley.

Woodworth, R. S. (1938). *Experimental Psychology*. New York: Holt.

Woodworth, R. S., & Schlosberg, H. (1954). *Experimental Psychology* (rev. ed.) (pp. 90–105). New York: Holt.

Wundt, W. *An Introduction to Psychology*. (1912). (R. Pintner, translator, from the second German edition). London: Allen & Unwin. (Reprinted 1924.)

Charles Chubb and George Sperling. Processing Stages in Non-Fourier Motion Perception. *Investigative Ophthalmology and Visual Science, 1988,* 29, No. 3, *ARVO Supplement,* 266.

PROCESSING STAGES IN NON-FOURIER MOTION PERCEPTION

Charles Chubb and George Sperling. New York University

Most recent motion-perception models propose detectors that are more or less sharply tuned to stimulus energy at various spatio-temporal frequencies.[1] However, it is easy to construct random stimuli which *do not* systemmatically excite such Fourier-energy analytic mechanisms and which nonetheless display strong, consistent apparent motion across independent realizations (Chubb & Sperling, ARVO, 1987). We show that two initial stages, a linear bandpass filter followed by a rectifier (absolute value, square) would suffice to expose the motion information carried by most nonFourier stimuli to subsequent Fourier-energy analysis. However, we further demonstrate apparently moving stimuli that would require two successive pairs of linear filtering and rectification stages in order to be sensed by a Fourier-energy analyzer.

Both the optimal spatial frequency and the sensitivity of the nonFourier mechanism are lower than those of the Fourier-energy mechanism. We use these differences to construct apparently moving stimuli that grossly violate scale invariance: from afar, they are seen moving in one direction by the Fourier mechanism; from close, they are seen moving in the opposite direction by the nonFourier mechanism.[2]

[1] van Santen, J. P. H & Sperling, G. *J. Opt Soc. Am A* 1985, 2, 300-321.

[2] Supported by AFOSR Life Sciences Directorate Grant 85-0364

Atch-36

George Sperling and Thomas R. Riedl. Summation and masking between spatial frequency bands in dynamic natural visual stimuli Investigative Opthalmology and Visual Science, 1988, 29, No. 3, *ARVO Supplement, 139*

SUMMATION AND MASKING BETWEEN SPATIAL FREQUENCY BANDS IN DYNAMIC NATURAL VISUAL STIMULI

George Sperling and Thomas R. Riedl, New York University

Dynamic images of a signer producing individual signs of American Sign Language (ASL) were bandpass filtered in adjacent spatial frequency bands. Intelligibility of a band was determined by testing deaf subjects fluent in ASL. By iteratively varying the center frequencies and bandwidths of the spatial bandpass filters, it was possible to divide the original signal (96 × 64 pixels) into four adjacent, intelligible, frequency bands with mean frequencies of 3.0, 7.5, 15, and 25 cycles per frame-width. All bands were found to have the same temporal frequency spectrum up to a multiplicative constant.

Masking of signals in band i by noise in band j (4×4 conditions) was measured by a rating method. The power ratio within a band i, $P_{signal}(i)/P_{noise}(i)$, required to produce a criterion rating response was the same for bands 2, 3, 4 and higher for band 1 (3 c/frame). The logarithm of the normalized crossband masking effectiveness was inversely proportional to $\log |(freq_{signal}/freq_{noise})|$. The 0.7 indicates an asymmetry: Masking of high frequency signals by low frequency noise is slightly greater the masking of low frequencies by highs.

Weak signals from bands i and j were linearly added and tested for intelligibility. Intelligibility was slightly greater for signals in the same band ($i = j$) versus adjacent bands, and for adjacent bands versus distant bands. Obviously, for *strong* signals, adding different bands produces more-intelligible combinations than does increasing power within a band.

The high intelligibility achieved by the narrow bands of our low resolution signals indicates that high-resolution broad-spectrum signals could be decomposed into many nonoverlapping frequency bands, each of which contained sufficient information for interpreting ASL.[1]

[1]Supported in part by AFOSR Life Sciences Directorate Grant 85-0364 and NSF Science and Technology to Aid the Handicapped, Grant PFR-80171189

Atch-27